FORUM

Pest Control by the Introduction of a Conditional Lethal Trait on Multiple Loci: Potential, Limitations, and Optimal Strategies

PAUL SCHLIEKELMAN[1] AND FRED GOULD[2]

J. Econ. Entomol. 93(6): 1543–1565 (2000)

ABSTRACT Advances in genetics have made it feasible to genetically engineer insect strains carrying a conditional lethal trait on multiple loci. We model the release into a target pest population of insects carrying a dominant and fully penetrant conditional lethal trait on 1–20 loci. Delaying the lethality for several generations after release allows the trait to become widely spread in the target population before being activated. To determine effectiveness and optimal strategies for such releases, we vary release size, number of generations until the conditional lethality, nonconditional fitness cost resulting from gene insertions, and fitness reduction associated with laboratory rearing. We show that conditional lethal releases are potentially orders of magnitude more effective than sterile male releases of equal size, and that far smaller release sizes may be required for this approach than necessary with sterile males. For example, a release of male insects carrying a conditional lethal allele that is activated in the F_4 generation on 10 loci reduces the target population to 10^{-4} of no-release size if there are initially two released males for every wild male. We show how the effectiveness of conditional lethal releases decreases as the nonconditional fitness reduction (i.e., fitness reduction before the trait becomes lethal) associated with the conditional lethal genes increases. For example, if there is a 5% nonconditional fitness cost per conditional lethal allele, then a 2:1 (released male:wild male) release with conditional lethal alleles that are activated in the F_4 generation reduces the population to 2–5% (depending on the degree of density dependence) of the no-release size. If there is a per-allele reduction in fitness, then as the number of loci is increased there is a trade-off between the fraction of offspring carrying at least one conditional lethal allele and the fitness of the released insects. We calculate the optimal number of loci on which to insert the conditional lethal gene given various conditions. In addition, we show how laboratory-rearing fitness costs, density-dependence, and all-male versus male-female releases affect the efficiency of conditional lethal releases.

KEY WORDS conditional lethal, genetic control, multilocus, sterile male, model

THE CONCEPT OF mass releasing genetically altered insects for purposes of pest control dates back more than 50 yr. Serebrovsky (1940) proposed the use of insects carrying chromosomal translocations, and Knipling (1955) proposed the sterile insect technique (SIT) for the control of insect pests. Although translocations have not been widely used in pest control, SIT has been applied extensively against a variety of insects. Although it has had many successes (Klassen et al. 1994), experience has shown that it works only when either there are resources available for a major wide-area effort (e.g., screwworm fly in the United States and Mexico) or it is applied to ecologically isolated areas. In either case, the biotic potential of the target population cannot be too high.

To achieve eradication, a high ratio of sterile to normal males must be achieved for multiple generations over an area large enough for migration to be low. To achieve high sterility in the target population, the ratio of sterilized to fertile insects must be high enough that a fertile insect has a low probability of mating with another fertile insect. If (1) the sterile insects in an all-male release have mating fitness equal to the wild males, (2) releases are done in such a way that the sterile insects are perfectly mixed with the wild ones, and (3) females mate only once, then the reduction in population in the next generation is equal to the percentage of sterile males in the male population. Unfortunately, the processes of sterilization and laboratory rearing result in individuals with fitness reduced up to 80% (e.g., Holbrook and Fujimoto 1970, Hooper and Katiyar 1971, Ohinata et al. 1971). Fitness in the field may be even worse (e.g., Shelly et al. 1994). Because of the low fitness of sterilized insects, very large ratios (on the order of 10:1–1,000:1) of sterile to wild insects are required to achieve a high level of sterility in the field. Because the sterilized insects are all dead after one generation and don't leave any descendants, releases must be continued regularly throughout the season to keep the ratio high.

In response to these problems, a number of alternate genetic control mechanisms have been proposed (see Whitten 1985), including natural sterility (e.g., hybrid sterility and cytoplasmic incompatibility).

[1] Biomathematics Graduate Program, Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203. Current address: Department of Integrative Biology, University of California, Berkeley, CA 94720-3140.

[2] Department of Entomology, North Carolina State University, Raleigh, NC 27695-7634.

translocations, meiotic drive and positive heterosis (to drive deleterious genes into the target population), and conditional lethal traits. Although these ideas have been around since the 1950s and 1960s, none have been implemented on a large scale, largely because of the difficulties in breeding and raising insects with the required characteristics. As a result, interest in these techniques declined in the 1980s and early 1990s.

The conditional lethal release method was first proposed by LaChance and Knipling (1962). In a conditional lethal release, insects are released into the field carrying a trait that is lethal only under restrictive conditions. If the trait does not become lethal immediately, then the trait can spread into the wild population through inter-mating between released and wild insects. Temperature-dependent lethal traits or traits causing a failure to diapause would be suitable.

Theoretical work on the use of conditional lethal traits for pest control has been relatively sparse. Klassen et al. (1970a) and Klassen et al. (1970b) modeled releases of insects carrying conditional lethal traits controlled by up to four genes with additive effects. They showed that such releases could achieve greater reduction in pest population numbers than sterile male releases. More recently, Kerremans and Franz (1995) modeled the release of females carrying a recessive temperature-sensitive lethal mutation and a Y-autosome translocation. They found that the use of a single conditional lethal gene would be ineffective, but that a single conditional lethal could be useful in combination with other techniques.

Primarily because of the lack of suitable genes, conditional lethal have not been used successfully for pest control. The production of dominant conditional lethal mutants by traditional genetic methods appears prohibitively difficult (Fryxell and Miller 1995).

Recent success in genetically transforming insects (Handler et al. 1998, Coates et al. 1998, Jasinskiene et al. 1998) raises the hope that the stable genetic transformation of insects may become a routine undertaking in the near future (Atkinson and O'Brochta 1999 review the progress of research in this area). Refinement of such genetic transformation techniques should allow the creation of insect strains carrying dominant conditional lethal traits. Conditional lethal genes known in, for example, *D. melanogaster* could be used with other species. Fryxell and Miller (1995) discussed the properties required of a conditional lethal gene suitable for general use as pest control agent and demonstrated the existence of one such gene in *D. melanogaster*.

Given these advances, it is appropriate to reexamine the use of conditional lethal traits for insect pest control. Conditional lethal releases with genetically transformed insects have (at least) three important differences from the conditional lethal releases envisioned by Klassen and coworkers. First, it should be possible to insert a much larger number of conditional lethal loci than could be attained through classical techniques. Second, nonconditional (i.e., constitutive) genetic load associated with the insertions is likely to be an important component of the dynamics of releases with conditional lethals on multiple loci. Third, Klassen focused primarily on cases in which the conditional lethal alleles are not fully penetrant. With genetic transformation techniques it should be possible to use conditional lethal alleles which can cause close to 100% mortality with a single copy. We use a genetic model to determine the optimal number of copies of a conditional lethal allele to engineer into a release strain for varied genetic and ecological conditions.

Specific Questions Addressed. In this article we assume that there is one dominant conditional lethal allele (i.e., coding DNA sequence) that can be inserted into multiple sites of an insect's genome. We assume that it would be feasible to insert copies of the allele onto between one and 20 loci that are not physically linked. We address the following specific questions:

Ideally, How Effective Can Multilocus Conditional Lethal Releases Be? How would the effectiveness of an ideal (i.e., released insects have no reduction in fitness) multilocus conditional lethal release compare with an ideal sterile male release? For a given size of release, what is the greatest reduction in the pest population that could be achieved?

What are the effects of a Fitness Cost Due To the Released Alleles? It is unlikely that a large number of gene copies can be inserted into the genome of an insect without doing nonconditional genetic damage (e.g., insertions within coding regions). Given this, how much nonconditional genetic load can the released insects carry and still be effective in spreading the conditional lethal gene into the pest population?

What is the Optimal Number of Loci at Which To Insert the Conditional Lethal Gene? The probability that the descendants of matings between released- and wild-type individuals will pass on at least one copy of the conditional lethal allele to their offspring is increased by each additional locus on which the conditional lethal allele is inserted. Obviously then, if the released alleles carry no fitness cost, then the eventual reduction in the pest population is greater for each additional locus. However, if each additional conditional lethal allele does carry a constituitively expressed fitness cost, then there should be an optimal locus number balancing the fitness reduction of the released insects and their offspring with the increase in fraction of offspring carrying the conditional lethal allele

How Much Difference in Effectiveness is there Between an All-Male Release and a Male-Female Release? How much is the pest potential of the population increased by release of females? A drawback to mass-release schemes for pest control is that the release can increase the number of pests in the field in the short run. In many cases, this is not economically acceptable. One way to overcome this is through male-only releases. Many agricultural pest species are pestiferous only in their larval stage. If only adult males are released, then there will be no increase in the number of larvae and no contribution to population growth. Unfortunately, separating insects by gender can be

difficult. Furthermore, releasing the females along with the males increases the size of the release and therefore increases the reduction in the pest population when the conditional lethal allele does become lethal. It would be useful to know how much is gained in long-term reduction and lost in short-term increase in pests when the females are released along with males.

How Much Do Decreases in Field Fitness Resulting from Laboratory Rearing Decrease the Effectiveness of This Technique? Insects raised in the laboratory often have a reduced fitness in field conditions. This can result from maternal effects, inbreeding, drift, or selection by laboratory conditions (see, for example Hopper et al. 1993 or Mackauer 1976). This laboratory-rearing load has been a major impediment to mass-release schemes. Laboratory-rearing effects with a genetic basis have stronger negative impacts in a conditional lethal release than a sterile male release because there are more generations of field selection against deleterious laboratory traits. It would be useful to know how much fitness reduction from specific laboratory-rearing traits can be sustained before conditional lethal releases become ineffective.

Methods

General Model Parameters and Terminology. The model simulates the release of genetically engineered insects of a diploid species into a wild population. The time scale of the model is generations. The model parameters and output variables are as given below:

Parameters. L: The number of loci on which the conditional lethal trait is inserted. *C:* The nonconditional fitness reduction per inserted conditional lethal allele. This is assumed to be caused by random damage to the genome caused by the insertion of the alleles. *I:* The fraction of the target population that are released insects immediately after the release.

Output Variables. (1) The frequency of the genotype with no conditional lethal alleles. Insects with this genotype will be the only survivors when the conditional lethal allele is activated and becomes lethal. (2) The number of insects with the no-conditional lethal genotype relative to the number in a population where there is no release, assuming no density dependence in mortality. This is the number (as opposed to frequency) of survivors when the conditional lethal allele becomes lethal. (3) The population size relative to that with no release, assuming no density dependence in mortality. This is calculated as the running product (over generations) of the average fitnesses. See below for more detail. (4) The gametic disequilibrium of the no-conditional lethal gamete type. This is defined as the difference between the no-conditional lethal gamete frequency and what it would be with no statistical association between loci.

Components of the Model. *Calculating Gametic and Genotype Frequencies.* In general, multilocus problems are difficult, because of the large number of genetic states possible with multiple loci (2^L possible gamete types and 3^L possible genotypes). Working with 10 loci requires 1,024 gamete types and 59,049 genotypes. Working with 20 loci requires over 1 million gamete types and over 3 billion genotypes. With the rapid advance of gene technology, a release involving 10–20 loci is conceivable. The multilocus problem is simple if the loci are in equilibrium. In this case the loci are in random association and gamete frequencies are simply the product of gene frequencies across loci. However, we are interested in the period immediately after the introduction of a population of one genotype into a population of another genotype. In this situation, there are correlations (henceforth called *gametic disequilibrium*) between allelic states at different loci. Initially, all individuals carrying the conditional lethal allele on one locus have it on all loci, and this association breaks down only gradually. Thus, we must track all genotypes as they change over time.

Fortunately, the problem can be simplified. The same trait is being released at the same frequency on every locus. If selection acts the same on all loci, then the frequencies of all genotypes with the same number of released alleles will be equal (regardless of which loci the alleles are on). In this case, we only have to track one variable for each possible number of released alleles per gamete: L variables instead of 3^L. The probabilistic calculations required to do this are involved, but they yield a fast and easily programmable algorithm (see *Appendix 1*). If we wish to break up the loci into multiple fitness types, then the number of variables is $(L_1+1)(L_2+1)(L_3+1)\ldots$, where L_1 is the number of loci with fitness type 1, and so on.

As inter-mating between released and wild populations occurs, the gametic disequilibrium breaks down. The gamete frequencies approach the products of the individual gene frequencies. In the absence of selection against the released trait, the gamete frequency of the no-conditional lethal gamete converges to $(1-p)^L$, where p is the initial frequency of the conditional lethal allele. The no-conditional lethal genotype frequency converges to the square of the gamete frequency: $(1-p)^{2L}$.

Nonconditional Genetic Load. The no-conditional lethal genotype insects (insects with no copies of the conditional lethal allele) are assumed to have maximum fitness. Other genotypes have a fitness reduction determined by the number of conditional lethal alleles. This fitness reduction is nonconditional: that is, it reduces the fitness of the insects carrying it under all environmental conditions, and therefore reduces the penetration of the conditional lethal allele into the target population. We normally assume that this fitness cost is the same on all loci. However, we also work with two classes of loci with distinct fitness costs to explore the sensitivity of results to the assumption of equal fitness costs.

Mackay et al. (1992) studied the effects of P-element insertions on viability of *Drosophila melanogaster*. They found that on average each insertion decreased viability of insects by 5.5% for heterozygotes and 12.2% for homozygotes (however, some insertions appeared to have no effect on viability). Regression of viability on the number of insertions

yielded an expression with significant linear and quadratic terms. The resulting quadratic expression is reasonably approximated by a multiplicative expression of the form $w(y) = (1\text{-}cost)^y$, where $w(y)$ is the fitness of the genotype, *cost* is the reduction in fitness per inserted allele, and y is the number of conditional lethal alleles in the genotype. Decreases in viability likely underestimate the total decrease in fitness. However, Mackay et al. were working with random insertions events; these insertions can be screened to find the ones that cause the least fitness damage. We use values of 2.5, 5, and 10% for *cost*. This expression applies until the conditional lethal allele becomes lethal, at which point the fitness of all individuals carrying any copies of the conditional lethal allele is zero.

When there are two types of loci, then marginal fitness is given by $w(y_1,y_2) = (1\text{-}cost1)^{y1}(1\text{-}cost2)^{y2}$, where $y1$ is the number of alleles with fitness cost *cost1* and $y2$ is the number of alleles with fitness cost *cost2*. See *Appendix 1* for more details on how selection is applied.

We assume throughout this article that mating is random and that there is no sexual selection. Thus, we cannot take into account (for example) preferences for mating between like types (e.g., wild with wild and released with released) or preferences of females of all genotypes for wild-type males. Such preferences can be crucial in sterile male releases. Thus, this assumption may be a major weakness in the model.

Population Dynamics. Numbers of surviving insects is the quantity of interest for pest control. Thus, genotype frequencies alone are not sufficient information. Although realistically modeling population dynamics is beyond the scope of this article, we can examine the two theoretical extremes of population regulation.

If mortality is density-independent, then the cumulative product of the population average fitnesses in each generation gives a measure of population size relative to what it would have been with no release. For example, assume a discrete population growth model with no density dependence and population growth rate $w_{avg}(t)R_t$, where $w_{avg}(t)$ is the average fitness at time relative to the fitness of a wild-type individual and R_t is the population growth rate at time t for a wild-type population. Then, the population at time $t + 1$ is

$$N_{t+1} = \bar{w}_t R_t N_t$$

and

$$N_t = \bar{w}_{t-1}R_{t-1}\bar{w}_{t-2}R_{t-2} \ldots R_0N_0 = \bar{w}_{t-1}\bar{w}_{t-2}\bar{w}_{t-3} \ldots \bar{w}_0R_tR_{t-1}R_{t-2} \ldots R_0N_0. \quad [1]$$

where t is the population size in generation 0 and N_t is the population size in generation t. If there is no release, then $w_{avg}(t) = 1$ for all t. If there is a release, then $w_{avg}(t)$ is less than 1 and the cumulative product of average fitnesses gives population size relative to that with no release. If this relative population size is multiplied with no-conditional lethal genotype frequency, then we have the number of no-conditional lethal insects (after a release) relative to the total number of insects with no release. This quantity will be referred to as the "relative number" of no-conditional lethal insects. It should be kept in mind that this quantity is meaningful only in the case of density independence.

The other extreme in population regulation occurs when density-dependence is so strong that the population is always at carrying capacity. In this case the no-conditional lethal genotype frequency alone is a measure of no-conditional lethal numbers relative to no release. Obviously, no real population behaves this way (and conditional lethal releases would not be useful against it if it did), but by looking at this extreme we can get a sense of the importance of population dynamics. Output of the population state will always show genotype frequencies and genotype numbers relative to no release.

Pest Damage. In the context of asking whether females should be included in conditional lethal releases, we are interested in how much additional pest damage will be caused by the release. This is linked with how the population's density is regulated. Again, we can only look at the extreme cases. If the population is always at carrying capacity, then population numbers remain constant and the release has no effect on population size and therefore no effect on pest damage. If mortality is density-independent, then the cumulative average fitness of the population is the population size relative to no release and is therefore related to pest damage. The average fitness of a generation is calculated as the average reproduction relative to that of a population with only wild-type insects. The relative number of eggs laid at the beginning of a generation is given by the cumulative average fitness through the previous generation, and the relative number of adults who reproduce in a generation is given by the cumulative average fitness through that generation. The pest damage caused by a generation should be proportional to a value between the previous generation's cumulative average fitness and the current generation's cumulative average fitness. If females are released, then cumulative average fitness becomes $w_{avg}(t\text{-}1)\ w_{avg}(t\text{-}2)\ w_{avg}(t\text{-}3)\cdots w_{avg}(0)\,(1+f)$, where f is the number females released relative to the number of wild females present before the release.

Laboratory-Rearing Costs. Fitness reduction of laboratory-reared insects can be grouped into maternal effects and genetic effects.

Maternal effects include malnutrition and diseases. Most maternal effects (excepting disease) will not extend beyond the release generation. Such effects are equivalent to a reduction in the size of the released population, and can be modeled as such. Diseases that can be transmitted vertically between generations are an exception to this. Such diseases may often be the primary agent for fitness reduction in laboratory populations (Hopper et al. 1993), but are beyond the scope of this article.

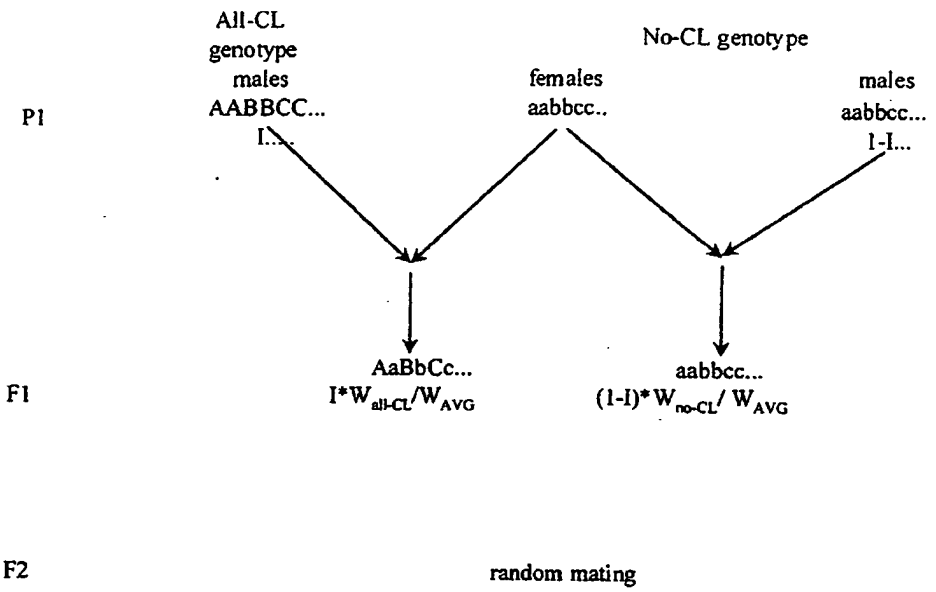


Fig. 1. Mating scheme for the all-male release. I is the fraction that the released males make up of the total male population. $W_{all\text{-}CL}$ is the fitness of the all-released genotype, $W_{no\text{-}CL}$ is the fitness of the no-conditional lethal genotype, and W_{avg} is the average fitness of the population.

Deleterious genetic changes in the laboratory can result from drift, inbreeding, or selection arising from laboratory conditions. Although deterioration in fitness of laboratory colonies has often been observed (e.g., Bigler et al. 1982, van Bergeijk et al. 1989, Geden et al. 1992), the genetic basis of such change has rarely been verified, let alone explored in detail (Hopper et al. 1993). We will, however, assume that there is genetically based reduction in fitness resulting from laboratory rearing. Because most deleterious mutations are recessive (Lynch and Walsh 1998), we will model deleterious genetic effects resulting from laboratory rearing as recessive traits.

Deleterious recessive traits are modeled as a group of J identical loci, such that an insect homozygous for the laboratory trait on all J loci experiences maximum fitness reduction. The relative fitness of a genotype with z loci homozygous for the laboratory trait and y copies of the conditional lethal allele is $w(x,y) = (1\text{-}cost)^y(1\text{-}labcost)^z$ where *cost* is the fitness reduction from each conditional lethal allele and *labcost* is the fitness reduction for each homozygous recessive locus.

It can be shown (Schliekelman 2000) that the number of loci contributing to a fixed total fitness reduction on the laboratory trait loci does not affect the frequency of insects that are no-conditional lethal on the conditional lethal loci. For simplicity, we use two loci for the laboratory trait and 50 and 20% (in separate model runs) as the relative fitnesses of individuals homozygous for the laboratory allele on both loci.

Release of Insects. Two types of insect releases are simulated: all-male and male-female. The size of releases is expressed as the ratio of number released to total (both genders) number of wild insects. Examples of both types of releases are given in *Appendix* 2.

In the all-male release there will be two types of matings in the release generation: no-conditional lethal genotype males with no-conditional lethal females and all-conditional lethal genotype males with no-conditional lethal females. See Fig. 1. If I is the fraction of males with the released genotype, then the proportion of matings involving a male of all-conditional lethal type will be the product of I and the fitness of the all-conditional lethal genotype. The offspring from these matings will be hemizygous for the conditional lethal allele on each locus (the term *hemizygous* is used instead of heterozygous to indicate that there is no alternate allele to the conditional lethal allele). Likewise, the proportion of matings involving no-conditional lethal males will be $(1\text{-}I)$ multiplied by the fitness of the no-conditional lethal genotype. The offspring from these matings will carry no copies of the conditional lethal allele. In the second generation and beyond, mating is random and proceeds according to the algorithm described in *Appendix 1.*

We assume that the released insects experience the full fitness cost due to the conditional lethal allele. Thus, the results will be a conservative estimate of effectiveness of the conditional lethal release method. because some of the fitness reduction is likely to occur in immature life stages.

In this article, we use an all-male release as the standard. We show (see *Results* section) that the difference between all-male and male-female releases is small if the size of the release is the same (that is, the number of males in the all-male release equals the number of males and females in the male-female release). Thus, all results can be applied to either type of release.

In the male-female release, there is random mating in the release generation. The initial cumulative av-

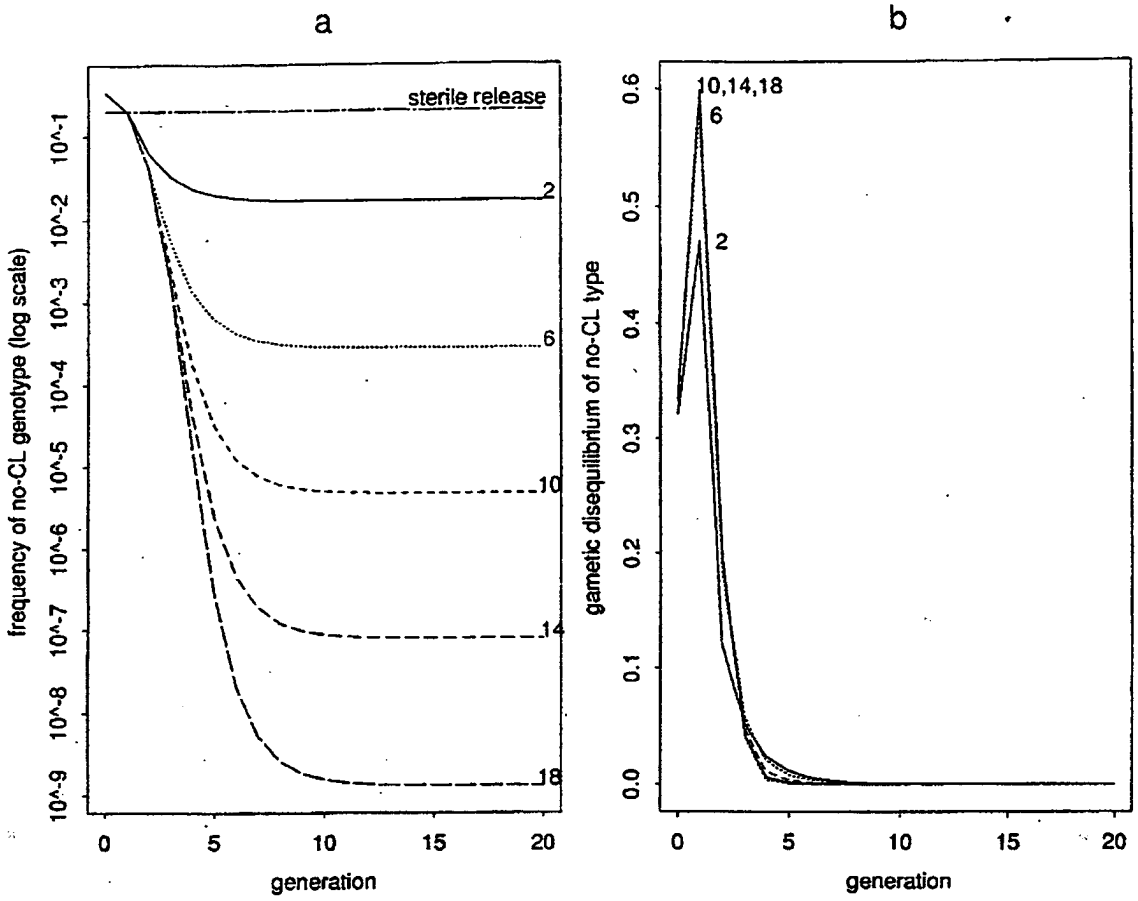


Fig. 2. 2:1 all-male release with no fitness cost for the released-type allele. (a) Frequency of the no-conditional lethal genotype shown on a log scale. (b) Gametic disequilibrium of the no-conditional lethal gamete type. This is calculated as $D = N_{AW} - p^L$, where N_{AW} is the frequency of the no-conditional lethal gamete and p is the no-conditional lethal allele frequency. See text for explanation of the shape.

erage fitness is adjusted to account for the increase in matings resulting from the release of females.

Sterile male and other genetic control releases have generally been very large. Typically, releases have been at a frequency of once per week or more (e.g., Davidson [1974]), with released-to-wild ratios of 100:1 not uncommon (e.g., Davidson [1974]). A conditional lethal release should require much smaller release sizes at less frequent intervals. We will simulate all-male release of sizes 1:6, 1:2, 2:1, and 19:2, which are chosen as a very small, small, medium, and large releases.

Results

"Ideal" Release. By *ideal* release we mean one in which the released insects have fitness equal to the wild type. This gives a bound on how much the method can reduce target population size, and is useful for comparisons with other genetic control techniques.

Figure 2a shows the frequency (on a log scale) of the no-conditional lethal genotype over time (generations) when an allele with no fitness cost is released on 2, 6, 10, 14, or 18 loci at a 2:1 ratio of released to native individuals. Only the no-conditional lethal genotype will survive when the conditional lethal allele becomes lethal.

The no-conditional lethal frequency drops rapidly with generation as the gametic disequilibrium breaks down and alleles from the released strain become dispersed throughout the wild population. As the disequilibrium drops, the no-conditional lethal frequencies approach their equilibrium values of 0.6^{2L} (where 0.6 is the frequency of the conditional lethal allele on each individual locus. See *Appendix* 2). Table 1 shows the no-conditional lethal genotype frequencies at two, four, and six generations as a function of L. By the fourth generation, the frequency of the no-conditional lethal genotype drops to under 4.5×10^{-4} for $L = 10$ and under 2.4×10^{-5} for $L = 18$. Table 1 also shows the same information for 1:6 and 19:2 all-male releases. Even small releases can be effective for large L. With a release as small as one released for every six in the wild population, the wild population can potentially be reduced to 12% of its original size if the conditional lethal allele is activated in the F_4 generation and 2% of its original size if the conditional lethal allele is acti-

Table 1. Fraction of no-CL genotype in the population in the F_2, F_4, and F_6 generations in a CL release with no fitness cost associated with the CL allele (fraction of remaining wild type insects after a SIT release is also shown)

Generation	Ratio of rel:wild	SIT	No. of loci in release				
			$L = 2$	$L = 6$	$L = 10$	$L = 14$	$L = 18$
F_1	1:6	0.75					
	2:1	0.20					
	19:2	0.05					
F_2	1:6		0.660	0.565	0.563	0.563	0.563
	2:1		0.160	0.045	0.040	0.040	0.040
	19:2		0.083	4.20×10^{-3}	2.59×10^{-3}	2.51×10^{-3}	2.51×10^{-3}
F_4	1:6		0.604	0.298	0.193	0.148	0.127
	2:1		0.137	5.07×10^{-3}	4.46×10^{-4}	7.88×10^{-5}	2.37×10^{-5}
	19:2		0.078	6.54×10^{-4}	1.06×10^{-5}	3.70×10^{-7}	2.72×10^{-7}
F_6	1:6		0.591	0.225	0.095	0.042	0.022
	2:1		0.131	2.69×10^{-3}	6.90×10^{-5}	2.23×10^{-6}	9.07×10^{-7}
	19:2		0.076	4.78×10^{-4}	3.32×10^{-6}	2.61×10^{-8}	2.36×10^{-10}

vated in the F_6 generation. The large (19:2) release reduces the no-conditional lethal population to frequencies of 10^{-7} to 10^{-10} in later generations.

The gametic disequilibrium of the no-conditional lethal type (the difference between the no-conditional lethal gamete frequency and that if loci were in equilibrium) is shown in Fig. 2b). The initial sharp increase is the result of a peculiarity of all-male releases explained in *Appendix* 2. The gametic disequilibrium drops sharply in the F_2 generation, but remains substantial for three to four generations. This indicates that a nonequilibrium model is necessary.

Effect of Nonconditional Fitness Cost of the Released Allele. *Five Percent Fitness Cost per Confidence Limits Allele.* Fig. 3 shows the no-conditional lethal genotype frequency (Fig. 3a) and relative number (Fig. 3b) in an 2:1 all-male introduction in which there is a 5% fitness cost per conditional lethal allele. In the release generation, all individuals either have no conditional lethal alleles or they have conditional lethal alleles in homozygous form at all loci. This second group of individuals have the minimum fitness. Accordingly, there is strong selection against the conditional lethal alleles in the release generation and the no-conditional lethal frequency increases. In the F_1 generation all individuals have either no conditional lethal alleles or they have conditional lethal alleles in hemizygous form at all loci (see Fig. 1). Thus, selection against the conditional lethal allele is weaker and the no-conditional lethal type drops in frequency in subsequent generations (except for $L = 16$–20) as inter-mating occurs between the native and released population and the gametic disequilibrium breaks down (and the number of individuals with many or no conditional lethal alleles decreases). However, the rate of disequilibrium breakdown decreases over time and selection is continuously acting in favor of the wild-type genes. Selection again becomes stronger than the impact of decreasing gametic disequilibrium after several generations and the no-conditional lethal genotype frequency begins increasing. Thus, there are three phases to the dynamics: phase 1, when the conditional lethal alleles are in high association, selection is dominant, and the frequency of no-conditional lethal genotype is increasing; phase 2, when the impact of the breakdown of gametic disequilibrium is stronger than selection and the no-conditional lethal genotype frequency decreases; and phase 3, when gametic disequilibrium has dissipated and selection against the conditional lethal allele becomes dominant again. All phases do not exist in all cases. The relative impact of selection and breakdown of gametic equilibrium (and thus the relative impact of the three phases) vary as L increases. For larger values of L there is very strong selection against the conditional lethal allele in the first few generations because all conditional lethal alleles are in association with many other conditional lethal alleles. With $L = 20$, for example, selection remains dominant through the F_2 generation and the no-conditional lethal genotype frequency increases rapidly. It is only when the conditional lethal alleles have become dispersed in the population (and thus the number of individuals with many conditional lethal alleles is low) that the effect of disequilibrium breakdown temporarily overpowers selection. Thus, the impact of phase 1 is great, whereas the impact of phase 2 is small. However, there is no phase 1 with $L = 2$. With few loci, selection against the conditional lethal allele is not strong enough to overcome the impact of the breakdown of disequilibrium until phase 3. This makes it clear why using more loci in a release is not always better.

Figure 3d (gametic disequilibrium of the no-conditional lethal gamete type) helps clarify this. The immediate strong selection against the conditional lethal allele in the high L case slows the rate of breakdown of gametic disequilibrium relative to lower L. Thus, lower L is favored if conditional lethal becomes lethal early. The efficacy of the release is sensitive to the generation in which conditional lethal becomes lethal. If the number of generations is small, then gametic disequilibrium (and, therefore, no-conditional lethal frequency) is still high. If the number of generations is too large, then selection in favor of the wild-type allele has driven no-conditional lethal frequency back-up.

Figure 3c shows the population size relative to no release in the case of density-independent mortality. We see that nonconditional genetic load causes sub-

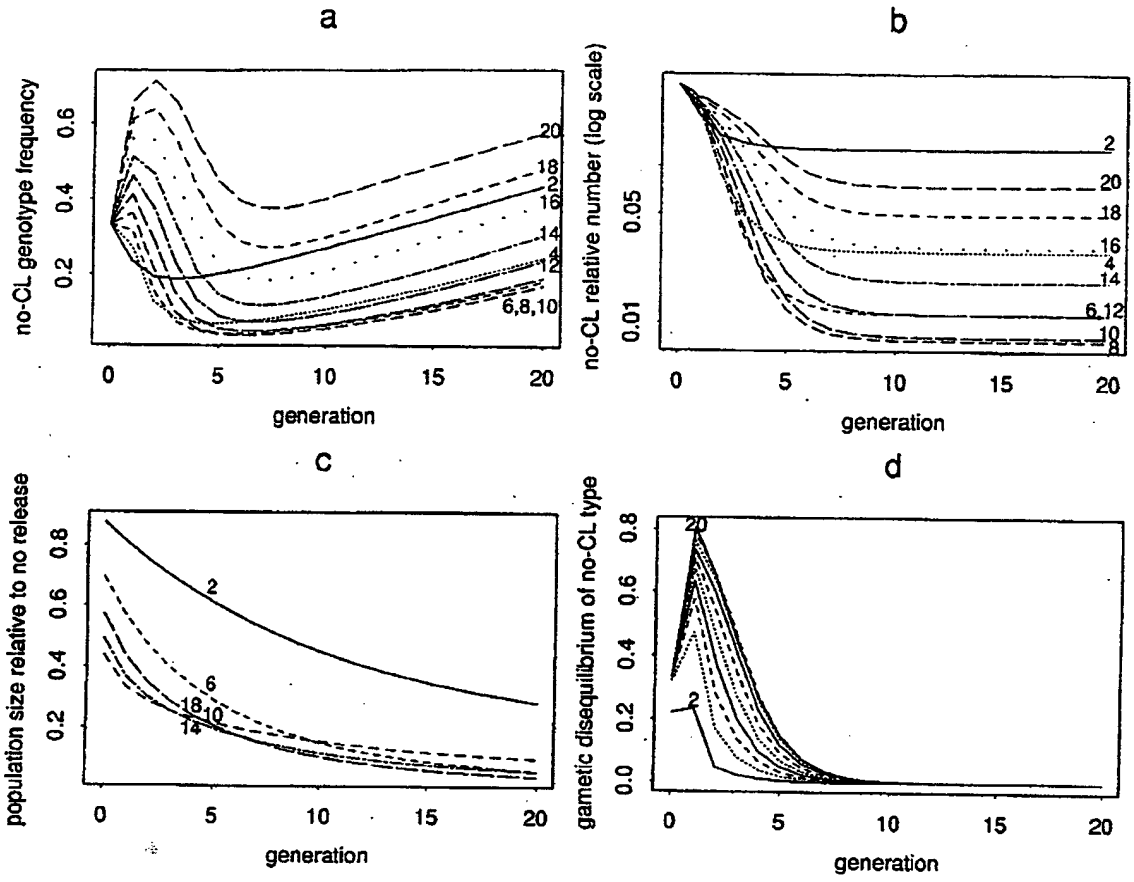


Fig. 3. 2:1 all-male release with a 5% cost per released-type allele in the genotype. (a) Frequency of no-conditional lethal genotype. In the case of population always at carrying capacity, this is the number of no-conditional lethal relative to that with no release. (b) Relative number of no-conditional lethal genotype. (No-conditional lethal genotype frequency multiplied by cumulative average fitness.) In the case of density-independent mortality, this is the number of no-conditional lethal genotype insects relative to no release. (c) Population size relative to no release when mortality is density independent. (d) Gametic disequilibrium of the no-conditional lethal gamete type. The curves go consecutively from $L = 2$ to $L = 20$ in the order shown.

stantial population suppression even before the conditional lethal alleles become lethal.

Note that the no-conditional lethal relative number reaches an asymptote as the gametic disequilibrium goes to zero. It can be shown that this occurs as a result of the fitness scheme used here (specifically, it requires that the fitness of a conditional lethal homozygote on one locus is equal to the square of the fitness of the hemizygote). Thus, the no-conditional lethal subpopulation grows at a rate that is independent of the genetic structure of the remainder of the population when gametic disequilibrium is zero. Note that the no-conditional lethal relative number will reach an asymptote with any fitness function as the average fitness goes to one and the conditional lethal alleles are removed from the population by selection (as opposed to the case here where it reaches an asymptote as the gametic disequilibrium dissipates).

The optimal number of loci to be used in a conditional lethal release can be determined from Fig. 3 a and b. In the case of a population always at carrying capacity, the optimal locus number is the value of L which minimizes no-conditional lethal genotype frequency in the generation that the released trait is expected to become lethal. In the case of density independent mortality, the optimal locus number is the L that minimizes the relative number of no-conditional lethal genotype individuals in the generation that the conditional lethal allele is activated. Optimal L depends strongly on the generation that the conditional lethal allele will become lethal—in the constant population size case the optimal L increases from 4 if the lethality occurs in the second generation to 6 if conditional lethal is lethal in the fourth generation and to 8 if the lethality occurs in the sixth generation. In the density-independent case, the optimal values are 4, 8, and 8, respectively. Note that these values are only accurate to ± 1, because only even values of L were checked.

At the optimal L, the no-conditional lethal numbers relative to populations with no release are significantly greater than in the ideal case, but still low: 0.05 in genotype frequency and 0.019 in relative numbers by the F_4 generation, and 0.04 and 0.0096, respectively, in

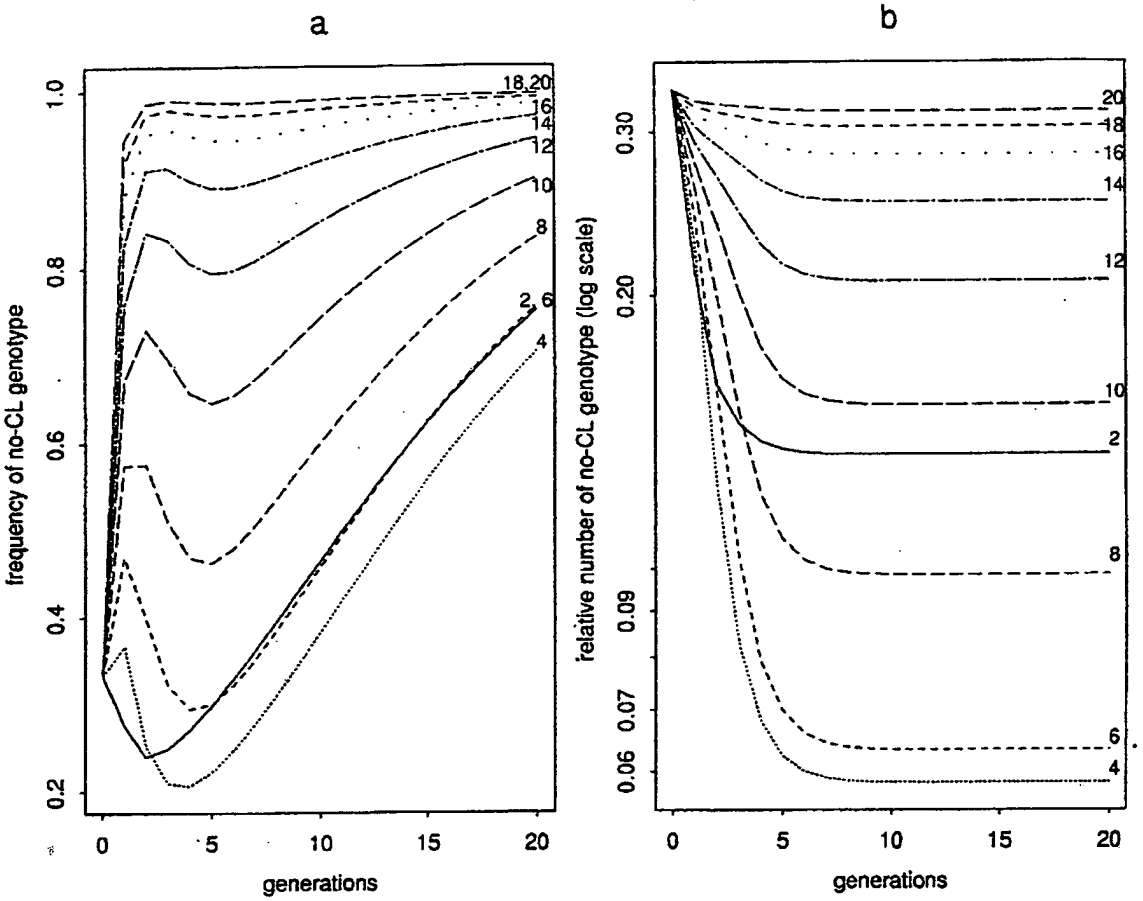


Fig. 4. 2:1 all-male release with a 10% cost per released-type allele in the genotype. (a) Frequency of no-conditional lethal genotype. In the case of population always at carrying capacity, this is the number of no-conditional lethal relative to that with no release. (b) Relative number of no-conditional lethal genotype. (No-conditional lethal genotype frequency multiplied by cumulative average fitness.) In the case of density-independent mortality, this is the number of no-conditional lethal genotype insects relative to no release.

the F_6 generation (with standard model conditions and a release size of 2:1).

Ten Percent Fitness Cost per Released Allele. Fig. 4 a and b shows the no-conditional lethal genotype frequency and relative number for a 2:1 all-male release with a 10% fitness cost per conditional lethal allele. The effect of the increased selection against the conditional lethal allele is obvious: selection against the conditional lethal alleles in phase 1 is very strong and the dispersal of the conditional lethal alleles in phase 2 causes little or no (depending on L) decrease in no-conditional lethal frequency before selection in favor of the no-conditional lethal type is dominants again in phase 3. For high L (16–20), the conditional lethal allele is almost completely removed from the population in the release generation. The efficacy of the release is diminished, with the lowest no-conditional lethal relative number attained being ≈0.060 (compared with 0.019 with a 5% cost per allele). The lowest no-conditional lethal frequency attained (for $L = 4$ with conditional lethal being lethal in the fourth generation) is 0.21 (compared with ≈0.05 with a 5% cost per allele). If mortality is density-independent, then the low average fitness of the population and the resulting decrease in total population size somewhat ameliorates the effect of the low competitiveness of insects carrying conditional lethal alleles.

The increased cost of the conditional lethal allele causes the optimal L values to be lower. For constant population size the optimal L is 2 if the conditional lethal allele becomes lethal in the F_2 generation and 4 if the conditional lethal allele becomes lethal later. When mortality is density-independent, the optimal value of L is 4 regardless of the generation that the conditional lethal allele becomes lethal.

Effect of Size of Release. *1:2 All-Male Release.* The previous results were for releases of a size which gives an initial 2:1 ratio of released to wild individuals. Fig. 5 a and b shows no-conditional lethal genotype frequency and relative number in a 1:2 all-male introduction with a 5% cost per released allele. The release is ineffective if the conditional lethal allele becomes lethal in the F_2 generation. This size of release can still be effective if the conditional lethal becomes lethal in the fourth or sixth generation, especially if the mortality is density independent (see Fig. 5b).

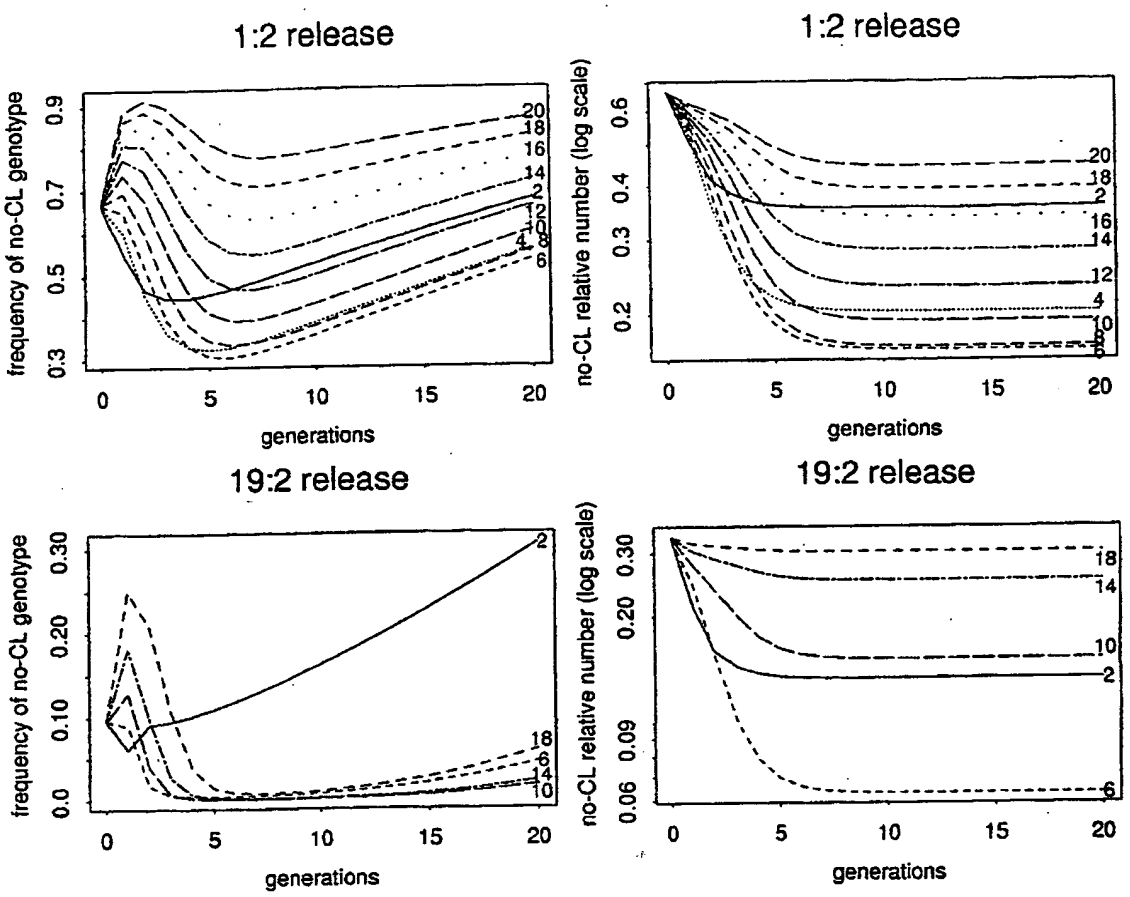


Fig. 5. Effect of size of release. Frequency and relative number of no-conditional lethal genotype insects in a 19:2 and 1:2 release.

19:2 All-Male Introduction. Fig. 5 c and d shows no-conditional lethal genotype frequency and relative number for a 19:2 all-male release. A large release with high L reduces the no-conditional lethal frequency to levels in the range of 10^{-7} and lower.

Comparing Figs. 3a, 5a, and 5c and 3a, 5b, and 5c, we see that the optimal L increases with the size of the release. This outcome is seen because the effectiveness of selection against the conditional lethal allele decreases as the no-conditional lethal genotype becomes rare when release ratios are high. With low frequency of the no-conditional lethal genotype, there is little variation in fitness for selection to act on.

Summary of the Effect of Conditional Lethal Allele Cost and Size of Release. Tables 2–4 show the (1) optimal L for the case of a population held at carrying capacity, (2) no-conditional lethal genotype frequencies at that L, (3) optimal L when mortality is density-independent, and (4) relative number of no-conditional lethal insects at the density-independent optimal L. Optimal L values are calculated for the conditional lethal allele activation in the F_2 (Table 2),

Table 2. Optimal locus numbers and frequencies when the CL allele becomes lethal in the F_2 generation

Cost	Ratio of released:wild	Optima L for constant population size	Frequency of no-CL genotype at optimal L	Optimal L (density-independent)	Relative no. of no-CL insects at optimal L
0.025	1:2	4	0.362	4	0.325
	2:1	6	0.086	6	0.064
	19:2	6	8.30×10^{-3}	8	4.87×10^{-3}
0.05	1:2	4	0.450	4	0.370
	2:1	4	0.129	4	0.087
	19:2	6	0.017	6	7.58×10^{-3}
0.10	1:2	2	0.547	2	0.448
	2:1	2	0.240	4	0.123
	19:2	4	0.044	4	0.015

Column 1: Optimal L when population size is always at carrying capacity. Column 2: Frequency in the F_2 generation at that optimal value of L. Column 3: Optimal L when mortality is density independent. Column 4: Relative number in the F_2 generation at that optimal value of L.

Table 3. Optimal locus numbers and frequencies when the CL allele becomes lethal in the F_4 generation

Cost	Ratio of released:wild	Optimal L for constant population size	Frequency of no-CL genotype at optimal L	Optimal L (density-independent)	Relative no. of no-CL insects at optimal L
0.025	1:2	8	0.153	10	0.110
	2:1	10	7.92×10^{-3}	12	3.35×10^{-3}
	19:2	14	3.94×10^{-3}	16	7.30×10^{-6}
0.05	1:2	4	0.330	6	0.218
	2:1	6	0.051	8	0.019
	19:2	10	1.72×10^{-3}	10	2.18×10^{-4}
0.10	1:2	2	0.568	4	0.363
	2:1	4	0.207	4	0.068
	19:2	6	0.034	6	0.003

Column 1: Optimal L when population size is always at carrying capacity. Column 2: Frequency in the F_4 generation at that optimal value of L. Column 3: Optimal L when mortality is density independent. Column 4: Relative number in the F_4 generation at the optimal value of L.

F_4 (Table 3), and F_6 (Table 4) generations for various combinations of release size and nonconditional fitness cost to the conditional lethal allele.

Comparison of All-Male and Male-Female Releases. Fig. 6 compares no-conditional lethal genotype frequency, no-conditional lethal relative number, and relative population size for four types of releases: (1) a 2:1 all-male release, (2) a 2:1 male-female release, (3) a 4:1 male-female release (all with a fitness cost of 5% for conditional lethal alleles), and (4) a 4:1 male-female release with a conditional lethal allele cost of 2.5%. L is set to 10 for all simulations of these releases. The total number of individuals in all of the 2:1 releases are equal and the total in all of the 4:1 releases are equal. The 4:1 male-female release has the same number of males as the 2:1 all-male release, but also has an equal number of females.

2:1 Releases. The no-conditional lethal genotype frequency initially increases more in the all-male release than in the equal sized male-female release (as explained in *Appendix* 2) After the F_1 generation, the separation between the all-male and male-female no-conditional lethal genotype frequency curves decreases, but there continues to be a higher portion of no-conditional lethal genotype insects in the all-male release indefinitely. However, there is very little difference between relative numbers of no-conditional lethal genotype insects in the two releases (Fig. 8b).

Equal Numbers of Released Males. The 2:1 all-male release and the 4:1 male-female release have the same number of males; the difference is in whether the females are also released. Fig. 8 a and b shows that releasing the females brings substantial benefit in reducing the frequency of no-conditional lethal types: there is a fivefold difference in no-conditional lethal genotype frequency between the two releases and a 15-fold difference in relative numbers of no-conditional lethal genotypes.

In the field in the male-female release, the initial relative population size (cumulative average fitness) reflects the increase in the number of females. Thus, it is initially (before selection is applied) set at 3 in the 2:1 release and 5 in the 4:1 release (Fig. 8c). Because $L = 10$, selection is strong against the conditional lethal allele. When mortality is density independent, population size drops quickly from the release generation level. The rate of the decrease is determined by the fitness cost of the released allele. If there is a 5% cost, then the population size drops below the nonrelease size (cumulative fitness = 1) in one generation. If the cost is 2.5%, then selection against the released insects is weaker, and the population size does not drop below the nonrelease size until the fourth generation (see Fig. 8c).

Two Types of Released Alleles. *Test of Sensitivity of Results to Equal Fitness Cost Assumption.* In the previous simulations it was assumed that the conditional lethal allele caused the same nonconditional decrease in the insect's fitness no matter where it was inserted within the genome. We use the two-allele type model to assess the importance of this assumption. We compare two releases: (1) a release with n

Table 4. Optimal locus numbers and frequencies when the CL allele becomes lethal in the F_6 generation

Cost	Ratio of released:wild	Optimal L for constant population size	Frequency of no-CL genotype at optimal L	Optimal L (density-independent)	Relative no. of no-CL insects at optimal L
0.025	1:2	10	0.0912	12	0.053
	2:1	14	1.58×10^{-3}	16	4.06×10^{-4}
	19:2	20+	6.71×10^{-7}	20+	4.24×10^{-8}
0.05	1:2	6	0.306	6	0.176
	2:1	8	0.0404	8	0.00956
	19:2	10	7.87×10^{-4}	12	3.86×10^{-5}
0.10	1:2	4	0.620	4	0.345
	2:1	4	0.247	4–6	0.0600
	19:2	6	0.0441	6	0.00214

Column 1: Optimal L when population size is always at carrying capacity. Column 2: Frequency in the F_6 generation at that optimal value of L. Column 3: Optimal L when mortality is density independent. Column 4: Relative number in the F_6 generation at that optimal value of L.

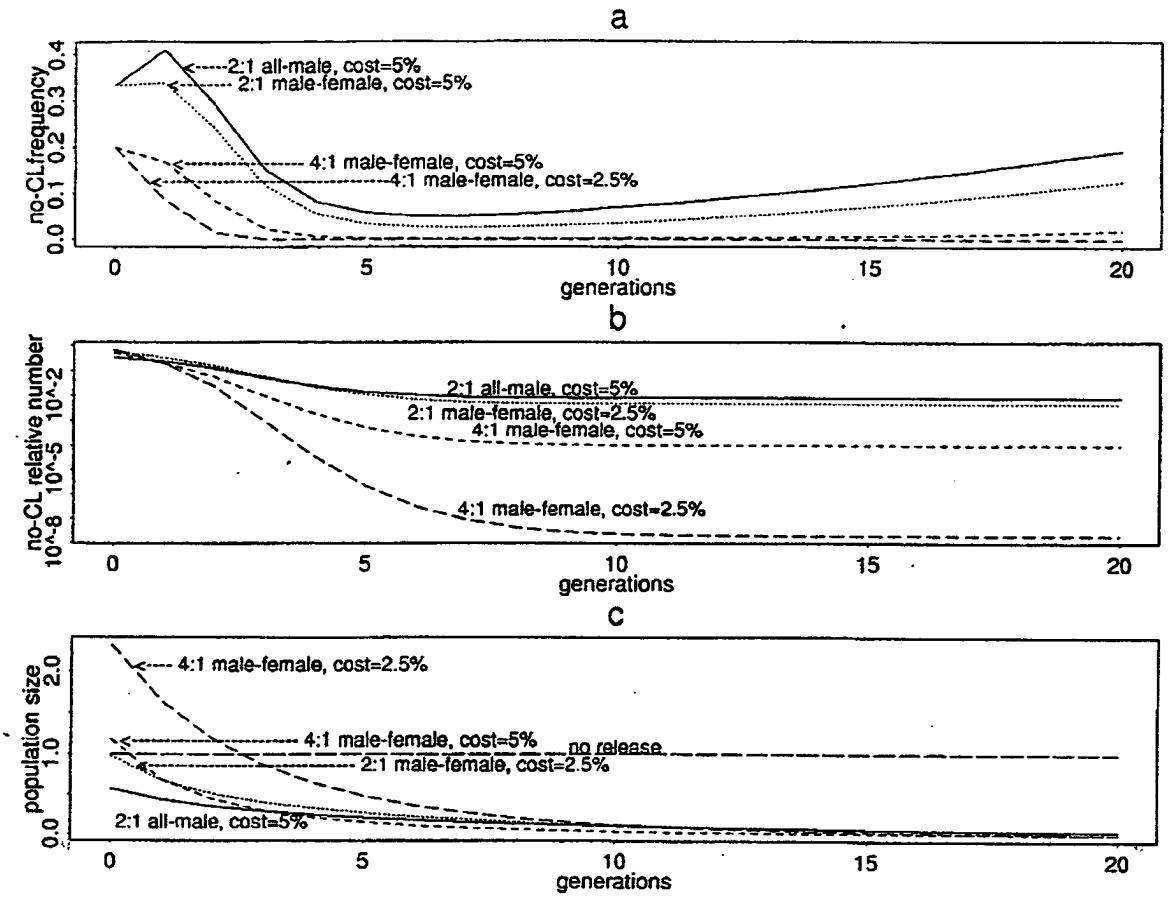


Fig. 6. Comparison of all-male and male-female releases. The 2:1 all-male and 2:1 male-female releases are releases of the same size, but with different gender distributions. The 4:1 male-female release has the same number of males as the 2:1 all-male release, but includes the females not released in the all-male release. (a) Frequency of no-conditional lethal genotype. In the case of complete density-dependence, this is the number of no-conditional lethal relative to that with no release. (b) Relative number of no-conditional lethal genotype insects. (All-wild genotype frequency multiplied by cumulative average fitness.) In the case of density-independent mortality, this is the number of no-conditional lethal genotype insects relative to no release. (c) Cumulative average fitness. If mortality is density-independent, then this is the population size relative to no release. In a release with females, this is initially set to account for the increase in breeding population size brought by releasing females (see *Methods*).

loci of fitness cost c_1 per allele and n loci of cost c_2 per allele, and (2) a release with $2n$ loci with fitness cost $1-\sqrt{(1-c_1)(1-c_2)}$. The cost function in (2) is midway between c_1 and c_2 and makes the all-conditional lethal genotype have the same fitness in both cases. Thus, we test how well a release with two allele types is approximated by a release with one allele type of the average effect. Fig. 7 shows this comparison for $L = 5$ with allele costs of $c_1 = 0\%$ and $c_2 = 10\%$ versus allele cost of $c_1 = c_2 = 0.0513$ (as given by the above formula). The no-conditional lethal frequency is indistinguishable between the two releases until at least the fifth generation, whereas the no-conditional lethal relative number is indistinguishable indefinitely. Given that few insect species of interest have more than approximately six generations per year, this indicates that the assumption of equal fitness cost of conditional lethal alleles across loci has negligible effect on the results.

Optimal Locus Number for Two Allele Types. We can also explore effectiveness and optimal locus number for a release with two different allele costs. A typical situation might be that the conditional lethal allele has been successfully inserted at low fitness cost on a small number of loci and at higher fitness cost on other loci. It would then be useful to know how many of the higher cost loci to use in the release. Fig. 8 a and b shows no-conditional lethal genotype frequency and relative number for a 2:1 all-male release with a 0 cost conditional lethal allele on three loci and 0.05 cost conditional lethal allele on L_2 loci, where L_2 varies from 0 to 14. This release is very effective, achieving reductions of 98.5% in frequency and 99.2% in relative numbers of no-conditional lethal types by the fourth generation for optimal L. The no-conditional lethal genotype frequencies increase only slowly from the minimum brought about by the breakdown of gametic disequilibrium. Table 5 shows optimal values of L_2.

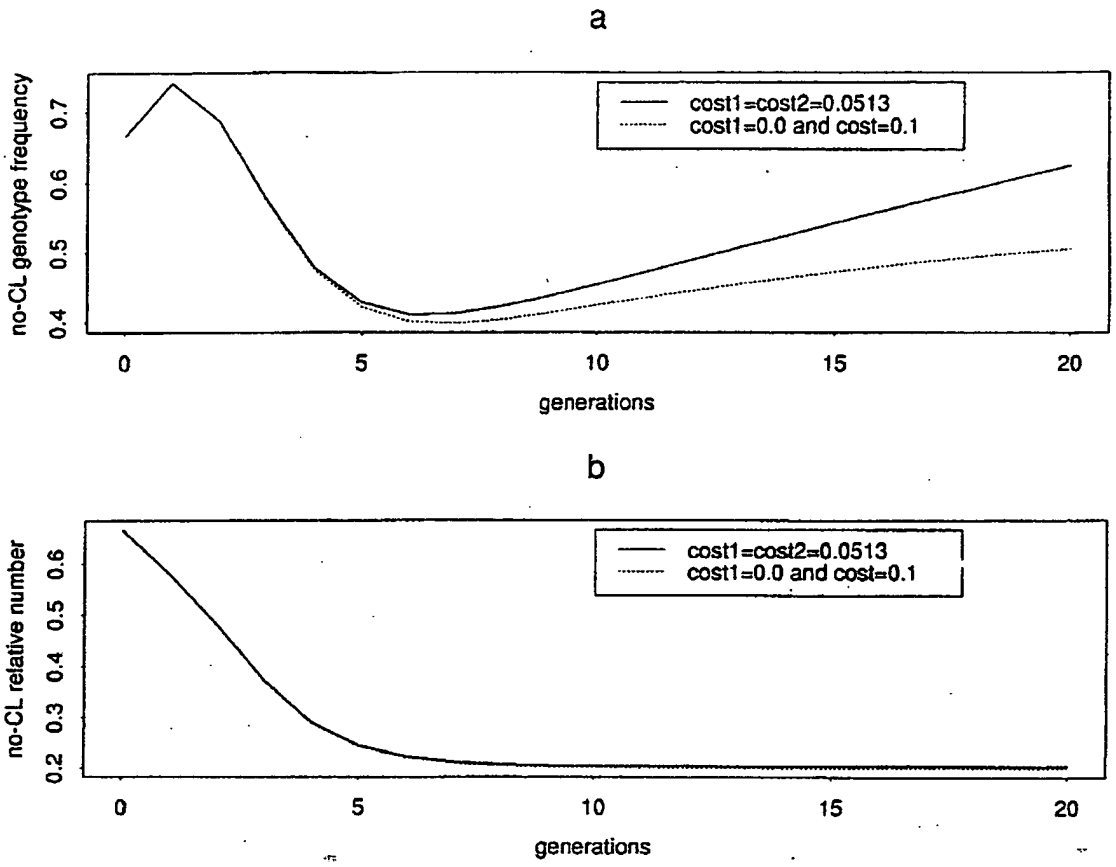


Fig. 7. Comparison of a release with 10 alleles of fitness cost 0.0513% each and a release with five alleles of 0 fitness cost and five alleles with 10% fitness cost each. These numbers were chosen to make the fitness of the all-released type equal in both releases.

along with the frequencies and relative numbers of the no-conditional lethal genotype. The optimal L_2 is 0 when the conditional lethal becomes lethal in the F_2 generation, 4 in the F_4 generation, and 6 in the F_6 generation for both extremes of density dependence. It is better to only use the three 0 cost loci if the conditional lethal allele becomes lethal in the F_2 generation, but it becomes beneficial to use the higher cost loci if the conditional lethal allele becomes lethal in later generations.

Having even a few very low cost loci is very beneficial if the conditional lethal allele becomes lethal in later generations. Compare Table 5 with Tables 2–4, which show data for the same release without the three loci with zero cost conditional lethal alleles. The benefit given by the three 0 cost loci is small if the conditional lethal allele becomes lethal in the F_2 generation (phase 2). Because breakdown of gametic disequilibrium is the dominant force for change in genotype frequencies at this stage, decreasing the fitness cost on a few loci has little effect. The decrease in fitness cost does have an effect in later generations when selection becomes more important. The benefit of the 0 cost loci is substantial if the conditional lethal allele becomes lethal in the F_4 and F_6 generations, decreasing the number of surviving insects by factors of ≈2 and 4, respectively.

Figure 8 c and d shows the same release, but with a 0.025 cost conditional lethal allele replacing the 0 cost allele. Table 5 has optimal L_2 and frequencies at optimal L_2 for this release. Again, it is best to use only the loci with low fitness cost if the conditional lethal allele becomes lethal in the F_2 generation but use the higher cost loci if lethality occurs in later generations. Again, the addition of the loci with low fitness cost has little effect until later generations. It is not until the F_6 generation that the addition of the three 0.025 cost loci brings any substantial benefit over the release without these insertions.

Effect of Laboratory Trait Fitness Cost. The purpose of this section is to compare releases in which the insects carry deleterious recessive traits resulting from laboratory rearing to releases in which insects are free of such traits.

Figure 9 a and b shows a 2:1 release with a 0% cost conditional lethal allele on L loci and deleterious recessive laboratory traits on two loci. The fitness of individuals homozygous at both loci for the laboratory trait is set at 50% of the fitness for the same genotype without any laboratory trait alleles. For $L = 10$ in the

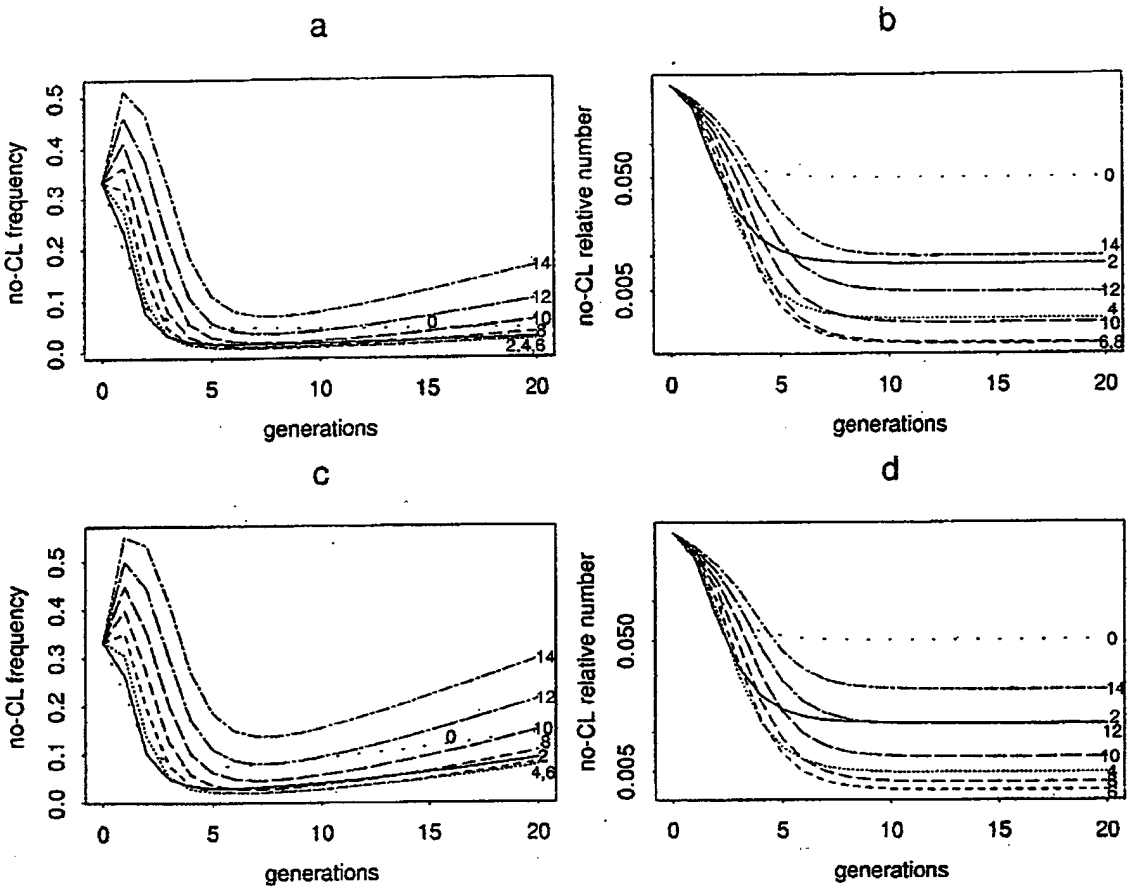


Fig. 8. Releases with a mix of higher and lower cost conditional lethal alleles. (a and c) A 2:1 release with insects carrying conditional lethal alleles with 0 fitness cost on three loci and 5% fitness cost on a varying number of loci. (c and d) A 2:1 release with insects carrying conditional lethal alleles with 2.5% fitness cost on three loci and 5% fitness cost on a varying number of loci (number shown by curve).

F_4 generation, the frequency of the genotype with no conditional lethal alleles (but laboratory trait alleles possible) is 4.7×10^{-3} and the relative number is 2.8×10^{-3}. These are an order of magnitude higher than for the same release with no laboratory trait costs (comparing with $L = 10$ in Table 1). In the F_6 generation the frequency and relative number are 9.8×10^{-4} and 5.2×10^{-4}, again an order of magnitude higher than the release with the no laboratory-associated traits.

Figure 9 c and d shows the same release, but with a fitness cost of 80% for individuals homozygous for the laboratory trait on both loci. The no-conditional lethal genotype frequency and number are ≈0.05 and 0.02 for $L = 10$ in the F_4 generation. In the F_6 generation these are 0.02 and 7×10^{-3}, respectively.

Table 6 compares the reduction in wild population for SIT and conditional lethal releases (with $L = 10$) with 50 and 80% fitness reductions caused by laboratory-associated traits. In the SIT release, the 50 and 80% reduction in fitness increases the surviving genotype population by factors of 1.6 and 2, respectively (comparing Tables 1 and 6). In the conditional lethal releases, the surviving population increases by an order of magnitude when fitness is reduced by 50% and two orders of magnitude when fitness is reduced by 80%. The impact of genetically based laboratory fitness

Table 5. Optimal locus numbers for a CL release with 3 loci with CL alleles that carry an unconditional fitness cost of 0% or 2.5% and L_2 loci with CL alleles that carry a 5% unconditional fitness cost

Cost	Ratio of released:wild	Optimal L_2 for constant population size	Frequency of no-CL genotype at optimal L_2	Optimal L_2 (density-independent)	Relative no. of no-CL insects at optimal L_2
0	F_2	0	0.065	2	0.062
	F_4	4	0.015	4	0.0078
	F_6	6	0.0084	6	0.0024
0.05	F_2	0	0.074	0	0.066
	F_4	4	0.035	4	0.012
	F_6	4	0.019	6	0.0050

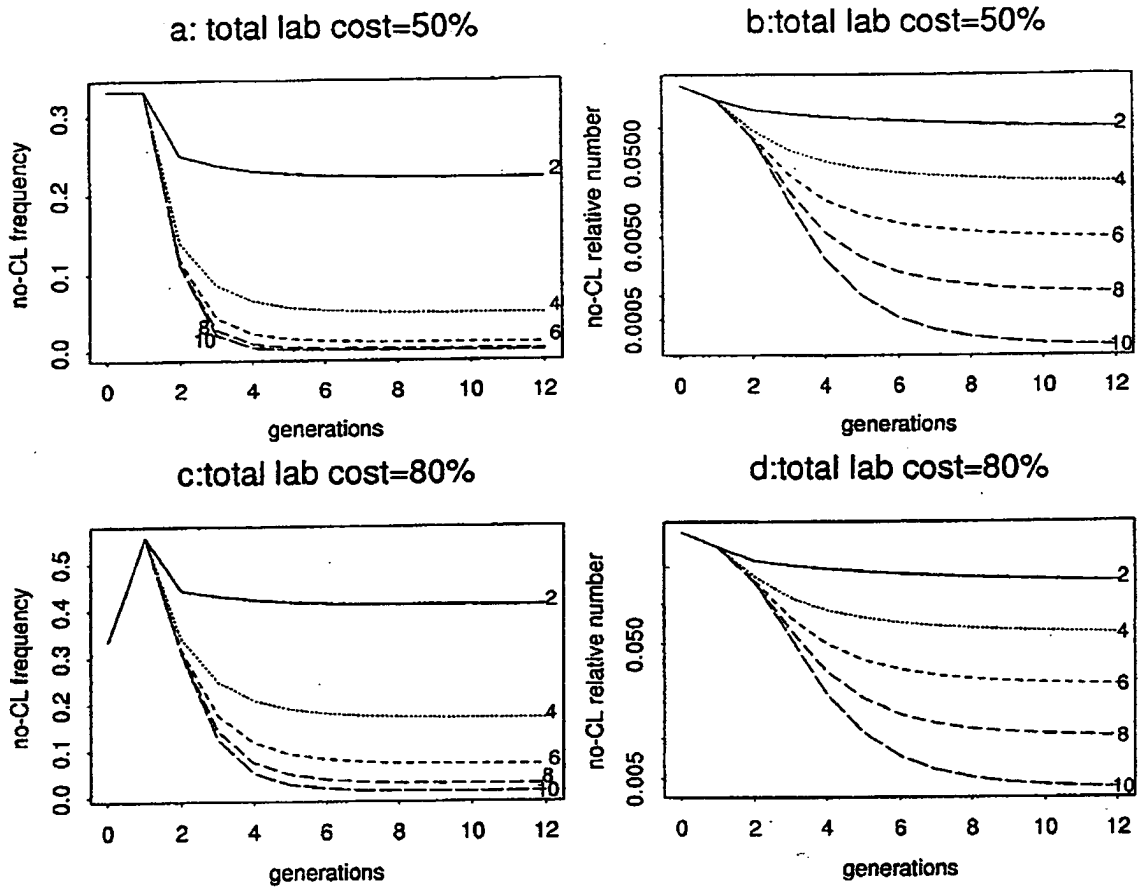


Fig. 9. Effect of fitness reduction associated with laboratory rearing. (a and b) 2:1 all-male release with a 0 fitness cost per released-type allele on L loci and 50% laboratory rearing fitness reduction determined by two recessive loci: The genotype has fitness 0.5 if both loci are homozygous for the laboratory trait allele, $\sqrt{0.5}$ if the genotype is homozygous for the laboratory trait on one locus, and one if neither. (c and d) 2:1 all-male release with a 0 fitness cost per released-type allele on L loci and 80% laboratory rearing fitness reduction determined by two recessive loci.

costs on conditional lethal releases is greater than on SIT releases.

Discussion

Ideally, How Effective Can Multilocus Conditional Lethal Releases Be? The potential effectiveness of the conditional lethal release method is best judged in comparison with the sterile male release method. Mass-release strategies suffer from many complications that we do not model here. Migration of insects, poor timing of release, weather conditions, and a multitude of other factors reduce the effectiveness of releases. Although we cannot model all of these factors, we can look at SIT, see how well it has done in various circumstances, and compare the likely effectiveness of conditional lethal releases with it.

An "ideal" sterile male release—a release where the sterile insects compete equally with the wild ones—will reduce the population in the next generation by a fraction equal to the fraction that the sterile males make of the total male population. For example, a 2:1

Table 6. Comparison of the effect of laboratory-associated fitness reductions on SIT release and on CL release

Generation	Ratio of released:wild	Relative frequency of surviving genotype			
		SIT (50% fitness reduction)	SIT (80% fitness reduction)	L = 10 (50% fitness reduction)	L = 10 (80% fitness reduction)
F_1	2:1	0.5	0.71		
F_2 CDD	2:1			0.112	0.309
DI				0.0743	0.144
F_4 CDD	2:1			4.67×10^{-3}	0.0507
DI				2.76×10^{-3}	0.0215
F_6 CDD	2:1			9.81×10^{-4}	0.0184
DI				5.20×10^{-4}	7.00×10^{-3}

SIT release reduces the wild population to 0.2 of its original size. By comparison, a 2:1 ideal conditional lethal release with $L = 18$ reduces the population to 0.04 of its original size if conditional lethal is lethal in the second generation, 0.000024 if conditional lethal is lethal in the fourth generation, and 9.1×10^{-8} if conditional lethal is lethal in the sixth generation. Conditional lethal releases are potentially a far more powerful technique than sterile insect releases. Even small releases of size 1:6 can be effective if a higher (4–6) number of generations is available for the conditional lethal alleles to spread before they become lethal.

Examination of Table 1 shows that the effectiveness of conditional lethal releases will not greatly exceed that of SIT for species with a small number (1–2) of generations before the conditional lethal is triggered: three or more generations are needed for the trait to spread into the wild population.

The modeling work of Klassen and coworkers (see *Introduction* for references) allowed up to three different conditional lethal traits, each controlled by up to four loci with additive effects. One of the cases studied in Klassen, Creech, and Bell (1970a), that of two or three conditional lethal traits controlled by a single autosomal locus each, is similar to what we have studied here. However, the potential for genetically engineering insect strains with a conditional lethal trait on many loci potentially makes the conditional lethal technique far more powerful than demonstrated by Klassen and coworkers. The work of Kerremans and Franz (1995) dealt with a conditional lethal gene on a single locus. They concluded that this was not effective compared with other strategies, but could be effective in combination with a Y-autosome translocation. However, even in combination with the translocation, this technique is much less effective than what we have demonstrated here—again, because genetic engineering techniques should allow many loci to be used.

What Effect Does a Fitness Cost of the Released Alleles Have? It is useful in comparing SIT and conditional lethal releases to separate fitness reductions caused by genetic manipulation from fitness reductions caused by laboratory rearing. Undesired selection by laboratory conditions is likely to be an inescapable feature of all strategies that use large-scale release of insects. The conditional lethal release method, and all mass-release strategies, will suffer from this problem. However, damage to the insect's genome caused by the insertion of conditional lethal alleles is, at least in principle, under the control of the geneticist. It may be possible to minimize such damage by improving techniques or by screening many insertion events for those with the least fitness cost.

The fitness reduction in SIT caused by irradiation (the "genetic manipulation" component) alone is on the order of 50–80% (see *Introduction* for references). This percentage translates directly to the reduction in effectiveness of SIT. The picture for conditional lethal releases is more complicated, because the fitness reduction depends on the number of insertions made into the insect's genome and because the selection occurs over multiple generations instead of just one.

It is clear that the effectiveness of conditional lethal releases decreases rapidly as the fitness cost of inserted alleles increases. Very small releases (e.g., 1:6 [see Table 1]) can reduce no-conditional lethal frequency to low levels when there is no fitness cost for inserted alleles. However, at a fitness cost of 5% per released allele, it takes a release on the order of 1:1 to get a useful reduction in no-conditional lethal frequency (see Tables 2–4). At a fitness cost of 10% per allele, it takes a release of size on the order of 10:1 (see Tables 2–4) to achieve meaningful reduction in no-conditional lethal frequency. If mortality is density-independent (and thus relative number is meaningful), the adverse effect of selection against released alleles is somewhat ameliorated by the increased genetic load incurred by the population as the cost of the released alleles increases. For example, examining Fig. 3c, we see that the population in the higher L cases is reduced to 30–40% of its original size by the F_4 generation without the conditional lethal allele ever becoming activated. Smaller releases might still be feasible at higher released allele cost if the density-dependence in mortality is weak.

The results from the model with two types of conditional lethal alleles (Fig. 8) indicate that if a small number of insertions can be made with zero or low fitness cost, the effectiveness of the conditional lethal release method is increased—even with high fitness costs of insertions on other loci. The zero cost alleles are selected against in the first few generations after the release, because of their genetic association with the higher cost released alleles. Once this association has broken down, however, the frequency of these alleles is largely untouched by selection and the corresponding loci contribute indefinitely to keeping no-conditional lethal genotype frequency low.

The results of Mackay et al. 1992 (see *Methods* section) indicate that an average fitness cost of 5% per insertion is attainable (in fact, better than 5% is attainable because insertion events can be screened and some insertion events in the Mackay et al. 1992 study appeared to have no negative effect on viability). Assuming that geneticists are never able to do better than this, how effective would conditional lethal releases be? From Tables 2–4, we see that a 2:1 all-male conditional lethal release with a 5% fitness cost per conditional lethal allele reduces the population to 1–12% of its original size, depending on the generation that the conditional lethal allele becomes lethal and on the degree of density dependence. By comparison, a 2:1 sterile male release with a 50% fitness of sterile males only reduces the population to 33% of its original size. If the fitness cost per insertion can be reduced to 2.5%, then the population can be reduced to well under 1% of its original size with one 2:1 release. In this case, conditional lethal releases are very powerful and it becomes feasible to do releases on a far smaller scale than the "overflooding" releases necessary with SIT.

If there is a fitness cost for the conditional lethal alleles, the timing of the lethality of the conditional

lethal allele becomes important. The high selection against the conditional lethal allele in the first few generations (when the conditional lethal alleles are in high association with each other) keeps no-conditional lethal frequency up and, thus, conditional lethal releases are not particularly effective if lethality occurs in the F_2 generation. Beyond the F_2, the breakdown of linkage disequilibrium overcomes this selection, and no-conditional lethal genotype frequencies drop rapidly. By about the seventh generation, selection predominates again and no-conditional lethal genotype frequencies rise again. If mortality is density-independent, no-conditional lethal numbers don't rebound. Although no-conditional lethal genotype frequencies are increasing, the total relative population size is actually decreasing (because population average fitness is less than one). If there is substantial density-dependence in mortality, then no-conditional lethal numbers rebound rapidly after the seventh generation.

What is the Optimal Number of Loci at Which To Introduce the Conditional Lethal Gene? Optimal number of loci is highly variable, depending on the fitness cost of conditional lethal alleles, the size of the release, and the generation in which the conditional lethal allele will become lethal. Optimal locus number decreases with increased fitness cost of the conditional lethal allele, increases with increasing size of the release, and increases with a higher number of generations before lethality. If lethality is in the second generation, the optimal locus number is in the range 4–6, except if the cost of the allele is very high (e.g., 10%). When lethality occurs in the F_4 and F_6 generations, the optimal L is also 4–6 when conditional lethal allele cost is 10%. For an allele cost of 5% with lethality in generations F_4 and F_6, optimal L is 4–8 for smaller releases (1:2 and 2:1) and 10 for the 19:2 release. The optimal L is most variable for an allele cost of 2.5%, ranging from 8 to 20 when lethality occurs in the F_4–F_6 generations. Optimal L is generally similar between the density-dependent and density-independent case, but occasionally higher in the density-independent case.

The picture becomes even more complicated when the fitness cost of the conditional lethal alleles varies between loci. However, comparisons of Table 5 with Tables 2–4 show that the optimal values for the total number of loci used are not much different between the examples with one allele type and two allele types (i.e., L_2+3 in Table 5 is close to the appropriate values of L in Tables 2–4). We do not know the extent to which this property applies in systems with more complicated variations in fitness cost.

How Much do Decreases in Field Fitness Caused By Laboratory Rearing Decrease the Effectiveness of This Technique? *Maternal/Environmental Effects.* Laboratory insects may have lower fitness because of direct environmental effects (e.g., crowded rearing conditions) or maternal effects. Because these effects are typically not heritable, they should not affect the conditional lethal approach much differently than they affect SIT.

Genetic Effects. Genetic load associated with laboratory strains decrease the effectiveness of the conditional lethal release method substantially. The no-conditional lethal genotype frequency increased by an order of magnitude for a 50% fitness reduction and by two orders of magnitude for an 80% reduction in fitness. These compare with increases on the order of two- to fivefold for SIT. Because of a longer period of field selection against the released insects, laboratory rearing fitness reductions have a greater effect on conditional lethal releases than SIT. For the 50% laboratory rearing cost, the conditional lethal releases still achieve a much larger reduction in population if the conditional lethal allele becomes lethal in later generations. When the laboratory rearing cost is 80%, the conditional lethal release method is still superior, but the margin is less than an order of magnitude.

Traditionally, the procedures for producing insects for mass release have emphasized quantity over quality. Given the theoretical potential for achieving major wild population reductions with much smaller released populations than with SIT, it may be better to try for smaller but genetically higher quality release populations for conditional lethal releases. The effectiveness of the two methods converges as quality of the released insects decreases.

How Much Difference in Effectiveness is There Between an All-Male Release and a Male-Female Release? *How is the Pest Potential of the Population Increased by Release of Females?* A laboratory population of insects cannot be raised without the females. Once it comes time for the release, there are two choices: devise some method for separating the females from the males or release the females along with the males and accept the increase in pest numbers. Separating females from males can be very costly, and is impossible in some cases (however, progress on genetic sexing systems has been made (Saul 1990, McCombs et al. 1993, McCombs and Saul 1995)).

In SIT, release of females is detrimental to control if the presence of sterile females decreases the number of matings between sterile males and wild females. If mating is random, this will occur only if the available number of female matings exceeds the available number of male matings. In a conditional lethal release, the release of females is beneficial to control as long as the mating is random, because it increases the frequency of the conditional lethal allele in the population. This is clear in a comparison of the 2:1 all-male and 4:1 male-female releases in Fig. 6. The no-conditional lethal genotype frequency in generations 4–6 when females are released is on the order of one-fifth the size of the no-conditional lethal frequency when females are not released. The difference in relative numbers of no-conditional lethal types is on the order of 10- to 20-fold. Thus, there is great benefit in releasing the females. This benefit is decreased if, due either to behavioral or spatial factors, released males are more likely to mate with released females. Unlike a sterile release, however, matings between released individuals do not completely "waste" the released insects because their descendants can still mate with wild

insects. However, if selection is strong against the all-conditional lethal genotype, such matings will increase the rate at which the released allele is removed from the population.

Releasing females would also be advantageous when the males in the target species are sexually selected. In such species, a small number of the most fit males get most of the available matings. Because released males are unlikely to be among the most fit males available, the fraction of matings involving released males is far less than that expected with random mating. Thus, the effectiveness of the release decreases if only males are released. However, it is beneficial if released females mate predominately with wild males. This minimizes the number of matings between high-conditional lethal genotypes, which are undesirable because of the low fitness of the offspring. Thus, release of females may be highly advantageous in species where females select male mates based on genetic fitness components. This aspect of conditional lethal releases may be of great benefit relative to sterile male releases.

Another aspect of mating biology with different impacts on SIT and conditional lethal releases is single versus multiple matings in females. SIT is less effective if females mate with multiple males. In conditional lethal releases, the number of matings per female has no impact on either all-male of male-female releases.

An equally important consideration is the increase in the size (and therefore pest potential) of the population caused by the release of females. If the fitness cost of the released-type alleles is high in a medium-sized release, then the genetic load introduced into the population quickly pulls population numbers below no-release levels. If the fitness cost of the released allele is low, then the population size can stay large for an extended period, and pest damage could be increased substantially (see Fig. 8c).

Evaluating the pest damage caused by a release requires more than the simple analysis here. We only wish to make the point that the genetic load brought by the released alleles can in itself suppress population levels if density dependence is not strong, and may be beneficial in short term population suppression.

Other Issues. Given the rapid progress of techniques for genetic transformation, it should be possible to mass release genetically engineered insects in the near future. We have shown that the release of insects carrying a dominant conditional lethal trait on multiple loci has the potential to be several orders of magnitude more effective than sterile insect releases. However, we have also shown that the effectiveness of conditional lethal releases is strongly dependent on the quality of the released insects. "Quality" can be divided into (at least) two parts: the quality of the genetic alterations and the genetic quality of the laboratory strains of the target species.

The effectiveness of conditional lethal releases decreases rapidly as the nonconditional fitness reduction caused by the insertions increases. If genetic insertion techniques never become better than causing an average 5% fitness reduction per insertion, then the effectiveness of conditional lethal releases is reduced to the order of a 5- to 10-fold superiority over SIT. However, if the conditional lethal trait can be introduced on even a few loci at a fitness cost close to zero, then the effectiveness increases.

Quality problems in mass rearing are also a serious problem. Because of a longer period of field selection, conditional lethal releases are more affected by genetic load resulting from laboratory rearing than SIT is. This effect alone can reduce the effectiveness of conditional lethal releases to within one to two orders of magnitude of SIT. Given the possibility of doing conditional lethal releases with much smaller and fewer releases than SIT, it may be feasible to rear insects of a higher quality than has been typical with SIT.

Given the assumptions that went into this modeling work, a large number of theoretical issues remain. The most important, perhaps, is nonrandom mating. As noted above, assortative mating behaviors can reduce the effectiveness of mass-release strategies. It is vital to understand the effect of such behaviors on the dynamics of the release. Spatial structure will also be important in species with very low or very high rates of dispersal. If dispersal is too low, then the released insects will remain in clumps and will not mate with wild insects. If dispersal is too high, then it will be difficult to maintain a sufficient density of conditional lethal-carrying insects in the target area. Dispersal rates will be a key factor in determining how large and often releases are needed with the conditional lethal release method.

Given that we did not find a great difference between the density independent and constant population size cases, population dynamics and age structure within a season may not be of great theoretical or practical importance in developing conditional lethal release strategies. However, population dynamics between seasons may be important. A great deal of theoretical work has been done on population dynamics and sterile male releases (see Ito et al. 1989 for a review), and most results should apply to conditional lethal releases.

Throughout this article we have assumed that the allele for conditional lethality is dominant and that 100% of individuals that have one or more copies of the allele will die once the conditional lethal trait is triggered. There is, therefore, an assumption that the phenomenon of gene silencing will not occur. Gene silencing occurs when individuals carrying multiple copies of a gene do not express the trait because of interference in the transcriptional or posttranscriptional process. The phenomenon of gene silencing has been documented in most detail for plants (Vaucheret et al. 1998, Grant 1999), but it has also been observed in mice (Garrick et al. 1998) and Drosophila (Pal-Bhadra et al. 1997, Birchler et al. 1999). In some studies of insects, no gene silencing has been noted at all (Spralding and Rubin 1983, Handler et al. 1998), and where it has been noted the "silencing" is more of a lowering of expression than it is a turning off of the gene (Pal-Bhadra et al. 1997). Therefore, lethal effects

could occur even with gene silencing if high levels of gene expression were not needed to interfere with survival. Additionally, studies of mice (Garrick at al. 1998) and Drosophila (Pal-Bhadra et al. 1997) indicate that the activation depends on the number of gene copies present in a specific generation. Therefore, as long as most of the insects have only one or two copies at the time that the conditional lethal is triggered, it will not matter how many copies are in each of the initially released insects.

In summary, conditional lethal releases will only be appropriate for certain insect species. New techniques for genetic manipulation of insects could greatly improve the efficiency of conditional lethal releases, but there will always be a need for close collaboration between molecular geneticists, insect ecologists, and pest management specialists if we are to gain the most benefit from this technology.

Acknowledgments

Thanks to Deborah Keys for assistance in the derivation of the multilocus algorithm. Thanks to Steve Ellner for insight and guidance throughout this project. Thanks to Sara Oppenheim, Amy Sheck, Nick Storer, Kyle Shertzer, Nikkala Pack, John Fieberg, Jonathon Rowell, and Juan Morales for discussion and review of the manuscript. This research was partially funded by a fellowship from the Agricultural and Life Sciences Department of North Carolina State University and by the Keck Foundation Progam in Behavioral Biology.

References Cited

Ashburner M., M. A. Hoy, and J. J. Peloquin. 1998. Prospects for the genetic transformation of arthropods. Insect Mol. Biol. 7: 201–213.

Atkinson, P. W., and D. A. O'Brochta. 1999. Genetic transformation of non-drosophilid insects by transposable elements. Ann. Entomol. Soc. Am. 92: 930–936.

Bigler F., J. Baldinger, and L. Luisoni. 1982. Impact of rearing method and host on intrinsic quality of *Trichogrammes evanescens*, pp. 167–180. *In* Les Colleques de l'INRA. No. 9. INRA, Antibes, France.

Birchler J. A., M. Pal-Bhadra, and U. Bhadra. 1999. Less from more: cosuppression of transposable elements. Nat. Genet. 21: 148–149.

Coates C. J., N. Jasinskiene, L. Miyashiro, and A. A. James. 1998. Mariner transposition and transformation of the yellow fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. U.S.A. 95: 3748–3751.

Davidson, G. 1974. Genetic control of insect pests. Academic, New York.

Fryxell K. J., and T. A. Miller. 1995. Autocidal biological control: a general strategy for insect control based on genetic transformation with highly conserved gen. J. Econ. Entomol. 88: 1221–1232.

Garrick D., S. Fiering, D.L.K. Martin, and E. Whitelaw. 1998. Repeat-induced gene silencing in mammals. Nature Genet. 18: 56–59.

Geden C.J., L. Smith, S. J. Long, and D. A. Rutz. 1992. Rapid deterioration of searching behavior, host destruction, and fecundity of the parasitoid *Muscidifurax raptor* (Hymenoptera: Pteromalidae) in culture. Ann. Entomol. Soc. Am. 85:179–187.

Grant, S. R. 1999. Dissecting the mechanisms of posttranscriptional gene silencing: divide and conquer. Cell 96: 303–306.

Handler A. M., S. D. McCombs, M. J. Fraser, S. H. Saul. 1998. The lepidopteran transposon vector, piggyBac, mediates germ-line transformation in the Mediterranean fruit fly. Proc. Natl. Acad. Sci. U.S.A. 94: 7520–7525.

Holbrook, F. R., and M. S. Fujimoto. 1970. Mating competitiveness of unirradiated and irradiated Mediterranean fruit flies. J. Econ. Entomol. 63: 1175–1176.

Hooper, G.H.S., and K. P. Kativar. 1971. Competitiveness of gamma-sterilized males of the Mediterranean fruit fly. J. Econ. Entomol. 64: 1068–1071.

Hopper K.R., R. T. Roush, and W. Powell. 1993. Management of genetics of biological-control introductions. Annu. Rev. Entomol. 38: 27–51.

Ito, Y., S. Miyai, and R. Hamada. 1989. Modeling systems in relation to control strategies, pp. 267–279. *In* A. S. Robinson and G. Hopper [eds.], Fruit flies their biology, natural enemies and control. Elsevier, Amsterdam.

Jasinskiene N., C. J. Coates, M. Q. Benedict, A. J. Cornel, C. S. Rafferty, A. A. James, and F. H. Collins. 1998. Stable, transposon-mediated transformation of the yellow fever mosquito, *Aedes aegypti*, using the Hermes element from the housefly. Proc. Natl. Acad. Sci. U.S.A. 95: 3743–3747.

Kerremans P., and G. Franz. 1995. Use of a temperature-sensitive lethal mutation strain of medfly (*Ceratitis capitata*) for the suppression of pest populations. Theor. Appl. Genet. 90: 511–518.

Klassen W., J. F. Creech, and R. A. Bell. 1970a. The potential for genetic suppression of insect populations by their adaptations to climate. U.S. Dep. Agric. ARS Misc. Publ. 11788.

Klassen W., E. F. Knipling, and J. U. Mcguire, Jr. 1970b. The potential for insect-population suppression by dominant conditional lethal traits. Ann. Entomol. Soc. Am. 63: 238–255.

Klassen W., D. A. Lindquist, and E. J. Buyckx. 1994. Overview of the joint FAO/IEA division's involvement in fruit fly sterile insects technique programs, pp. 3–26. *In* C. O. Calkins, W. Klassen, and P. Liedo [eds.], Fruit flies and the sterile insect technique. CRC, Boca Raton, FL.

Knipling, E. F. 1955. Possibilities of insect control or eradication through the use of sexually sterile males. J. Econ. Entomol. 48: 459–462.

LaChance L., and E. Knipling. 1962. Control of insects through genetic manipulations. Entomol. Soc. Am. Ann. 55: 515–520.

Lynch M., and B. Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer, Sunderland, MA.

Mackauer, M. 1976. Genetic problems in the production of biological control agents. Annu. Rev. Entomol. 21: 369–385.

Mackay T.F.C., R. F. Lyman, and M. S. Jackson. 1992. Effects of P element insertions on quantitative traits in *Drosophila melanogaster*. Genetics 130: 315–332.

McCombs S.D., S. G. Lee, and S. H. Saul. 1993. Translocation-based genetic sexing system to enhance the sterile insect technique against the melon fly. Ann. Entomol. Soc. Am. 86: 651–654.

McCombs S.D., and S. H. Saul. 1995. Translocation-based genetic sexing system for the oriental fruit fly (Diptera: tephritidae) based on the pupal color dimorpism. Ann. Entomol. Soc. Am. 88: 695–698.

Ohinata K., D. L. Chambers, M. Fujimoto, S. Kashiwai, and R. Miyabara. 1971. Sterilization of the Mediterranean fruit fly by irradiation: comparative mating effectiveness of treated pupae and adults. J. Econ. Entomol. 64: 781–784.

Pal-Bhadra M., U. Bhadra, and J. A. Birchler. 1997. Cosuppression in *Drosophila*: gene silencing of alcohol dehy-

drogenase by white-Adh transgenes is polycomb dependent. Cell 90: 479–490.

Saul, S. H. 1990. A genetic sexing system to improve the sterile insect technique against the Mediterranean fruit fly. J. Hered. 81: 75–78.

Schliekelman, P. 2000. Reassessing autocidal control. Ph.D. dissertation. North Carolina State University, Raleigh.

Serebrovsky, A. S. 1940. On the possibility of a new method for the control of insect pests. Originally published in 1940 in Zool. Zh. 19: 618–630 (English translation in Sterile male technique for the eradication of harmful insects).

Shelly, T. E., T. S. Whittier, and K. Y. Kaneshiro. 1994. Sterile insect release and the natural mating system of the Mediterranean fruit fly, *Ceratitis capitata* (Diptera: Tephritidae). Ann. Entomol. Soc. Am. 87: 470–481.

Spralding, A. C., and G. M. Rubin. 1983. The effect of chromsomal position on the expression of the *Drosophila xanthine* dehydrogenase gene. Cell 34: 47–57.

van Bergeijk K.E., F. Bigler, N. K. Kaashoek, and G. A. Pak. 1989. Changes in host acceptance and host suitability as an effect of rearing *Trichogramma maidis* on a factitious host. Entomol. Exp. Appl. 52: 229–238.

Vaucheret, H.,C. Beclin, T. Elmayan, F. Feuerbach, C. Godon, J.-B. Morel, P. Mourrain, J.-C. Palauqui, and S. Vernhettes. 1998. Transgene-induced gene silencing in plants. Plant J. 16: 651–659.

Whitten, M. J. 1985. The conceptual basis for genetic control. pp. 465–528. *In* G. A. Kerkut and L. I. Gilbert [eds.]. Comprehensive insect physiology, biochemistry, and pharmacology, vol. 12. Insect control. Pergamon. Oxford.

Received for publication 28 June 1999; accepted 7 June 2000.

Appendix 1. Derivation of the Algorithm for Multilocus Mating and Selection

The goal is to track the genotype frequencies when a laboratory population of insects carrying a conditional lethal trait at L loci is introduced into a wild population. We assume that the target species is diploid. It would be most advantageous to put each of the inserted genes on a different chromosome or linkage group because physical linkage would slow the rate at which the introduced genes penetrate the wild population. We therefore assume that all of the introduced loci are, in fact, unlinked (i.e., recombination frequency = 0.5 between all loci). Second, we assume that selection acts the same way on each of the loci. Under these two assumptions all genotypes and gamete types with the same number of introduced alleles will behave exactly the same way, and, in particular, will always be equal in frequency. This means that we just have to track one variable for each possible number of introduced alleles: L gamete types and L^2 genotypes. The set of all gametes with the same number of conditional lethal alleles will be referred to as a *gamete class.* The set of all genotypes with the same number of loci in homozygous form for the conditional lethal allele and the same number of loci with no conditional lethal alleles is called a genotype class.

We will write gamete types as vectors of 1's and 0's, where 1 denotes presence of the conditional lethal allele and 0 denotes absence of the conditional lethal allele. For example:

$$\begin{pmatrix}1\\0\\1\\0\end{pmatrix}$$

represents a gamete with four loci of concern that has a conditional lethal allele on the first and third loci. The loci will be numbered 1 to L starting from the top. Genotypes will be represented similarly:

$$\begin{pmatrix}1&0\\0&1\\1&1\\0&0\end{pmatrix}$$

Variables are defined as follows. L: number of loci on which the conditional lethal allele has been introduced. Y: gamete class. A scalar with the value being equal to the number of conditional lethal alleles in the gamete class. $P(Y)$: The probability that a randomly chosen gamete is of class Y. Y_1, Y_2: Parental gamete classes. These are scalars, with the value being equal to the number of conditional lethal alleles in the gamete class. $[Y_1, Y_2]$: parental gamete classes in a mating. $\vec{X} = [X_0, X_1]$: genotype class, where X_0 = the number of (0 0) pairs in the genotype (scalar), X_1 = the number of (1 1) pairs in the genotype (scalar). $P([X_0, X_1]:)$: the probability that a randomly chosen individual has genotype class $[X_0, X_1]$. $f([X_0, X_1])$: relative fitness of genotype relative to the genotype with maximum fitness.

$\bar{w} = \sum_{\vec{X}} P(\vec{X}) f(\vec{X})$: average fitness of the population.

$w(\vec{X}) = f(\vec{X}) / \bar{w}$: fitness of genotype $[X_0, X_1]$ relative to the population average fitness.

We want the distribution of the Ys (gamete classes) in the next generation given the distribution now. The frequency of a gamete class in the next generation is obtained by summing the probability of that gamete class being produced by a given mating over all possible matings in the current generation:

$$P(Y(t+1)) = \sum_{Y_1=1}^{L} \sum_{Y_2=1}^{L} P(Y(t+1) | Y_1(t), Y_2(t)) P([Y_1, Y_2]), \quad [A1]$$

Where $P(Y(t+1) | [Y_1(t), Y_2(t)])$ is the probability of gamete class Y resulting from a union between paren-

tal gametes of class Y_1 and Y_2, and $P(Y_1,Y_2)$ is the probability of a union between gametes of class Y_1 and Y_2 in the current generation.

To calculate $P(t+1)|[Y_1,Y_2]$ we must consider the distribution of genotype classes resulting from the union of gamete class Y_1 with gamete class Y_2. Thus, $P(Y(t+1)|[Y_1,Y_2]$ breaks down further:

$$P(Y(t+1)|Y_1(t), Y_2(t)) = \sum_{\vec{X}} P(Y|\vec{X})\,P(\vec{X}|[Y_1, Y_2]). \quad [A2]$$

$P[Y(t+1)]$ is then

$$P(Y(t+1)) = \sum_{Y_1=1}^{L}\sum_{Y_2=1}^{L}\sum_{\vec{X}} P(Y|\vec{X})\,P(\vec{X}|Y_1, Y_2)\,P([Y_1, Y_2]). \quad [A3]$$

We need to find expressions for the various parts of this equation. We can get $P(Y|[X_0,X_1])$ easily. This is the distribution of gametes produced by an individual of genotype class $[X_0,X_1]$. X_1 tells us the number of (1 1) pairs and X_0 tells us the number of (0 0) pairs in the genotype. All of the remaining loci have either a (1 0) or a (0 1). The resulting gamete will always have a 1-allele on loci with a (1 1) pair in the parental genotype and a 0-allele on loci with a (0 0) pair in the parental genotype. Because the recombination frequency is 1/2, then the number of 1-alleles arising (in the next generation gametes) from the (1 0) and (0 1) pairs are distributed as binomial (L-X_1-X_0, 1/2). L-X_1-X_0 is the number of (1 0) and (0 1) pairs. We then have

$$Y = X_1 + Z$$

where Z is distributed as a binomial distribution with L-X_1-X_0 trials with ½ probability of success.

Then

$$P(Y|[X_0, X_1]) = \begin{cases} 0 & \text{if } X_1 > Y \text{ or } X_0 > L - Y \\ P(Z = Y - X_1) & \text{else} \end{cases} \quad [A4]$$

Under an assumption of random mating, we have

$$P([Y_1, Y_2]) = P(Y_1)\,P(Y_2). \quad [A5]$$

Multiple matings per individual do not change anything provided that there is no sperm competition.

The last piece that we need is $P([X_0,X_1]|[Y_1,Y_2])$, the distribution of offspring genotype classes given the parental gamete classes. If we knew the specific parental gametes, then we would know the offspring genotype exactly. However, we only know the number of 1-alleles on each parental gamete. Knowing only this, there are usually multiple possibilities for the resulting offspring genotype. For example, if we have an introduction on three loci and pairing between a $Y = 2$ gamete and another $Y = 2$ gamete, there are a number of possibilities as to what exact mating is occurring:

$$\begin{pmatrix}0\\1\\1\end{pmatrix} \times \begin{pmatrix}0\\1\\1\end{pmatrix} \rightarrow \begin{pmatrix}0&0\\1&1\\1&1\end{pmatrix} \rightarrow \begin{pmatrix}0\\1\\1\end{pmatrix}$$

$$\begin{pmatrix}1\\0\\1\end{pmatrix} \times \begin{pmatrix}0\\1\\1\end{pmatrix} \rightarrow \begin{pmatrix}1&0\\0&1\\1&1\end{pmatrix} \rightarrow \begin{pmatrix}1 \text{ or } 0\\1 \text{ or } 0\\1\end{pmatrix}$$

$$\begin{pmatrix}1\\1\\0\end{pmatrix} \times \begin{pmatrix}0\\1\\1\end{pmatrix} \rightarrow \begin{pmatrix}1&0\\1&1\\0&1\end{pmatrix} \rightarrow \begin{pmatrix}1 \text{ or } 0\\1\\1 \text{ or } 0\end{pmatrix}$$

In the first mating the parental gametes are the same on all three loci. In this case the offspring gamete is determined from the parental gametes, because there is only one possible allele at each locus. In the second mating, the two gametes differ on the first two loci. Therefore, the offspring genotype is heterozygous on these two loci and gametes resulting from this genotype can have either 1 or 0 alleles on both loci. Because the genotype is homozygous (1 1) on the third locus. alleles on the third locus in gametes produced by this genotype are all 1-alleles.

Only certain values for X_0 and X_1 are possible for a given pair of parental gamete classes. We now determine what those are. Because Y_1 and Y_2 are interchangeable, we can assume $Y_1 < Y_2$ without loss of generality. For purposes of labeling, assume that Y_1 is the paternal gamete and Y_2 is the maternal gamete. This also causes no loss of generality. Assume there are L loci. We will start by determining the minimum values of X_0 and X_1 given Y_1 and Y_2. We will number the loci so that the Y_1 1-alleles on the paternal gamete are placed in locus positions 1 to Y_1. Now, the minimum number of matching pairs (0 0 or 1 1) between the gametes occurs if the maternal gamete (with Y_2 1-alleles) has its 1-alleles placed on locus L and the loci counting back from (i.e., loci L-Y_2 +1 to L). If $Y_1+Y_2 > L$, then there is an overlap between the 1-alleles on the two gametes, and there will be Y_1+Y_2-L pairs of 1-alleles. In this case there are no pairs of 0-alleles. If $Y_1+Y_2 < L$, then there is no overlap between 1-alleles and there are L-(Y_1+Y_2) pairs of 0-alleles. If $Y_1+Y_2 = L$, then there are no pairs of 1-alleles and no pairs of 0-alleles.

Example 1: Take $L = 6, Y_1 = 2$, and $Y_2 = 3$. If we arrange the 1-alleles as described to get the minimum number of pairs, we have

$$\begin{pmatrix}1\\1\\0\\0\\0\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}0\\0\\0\\1\\1\\1\end{pmatrix}$$

In this case there are no pairs of 1-alleles. but there are L-(Y_1+Y_2) = 6-(3+2) = one pair of 0-alleles.

Example 2: Now take $L = 6$, $Y_1 = 3$, and $Y_2 = 5$. We again arrange the 1-alleles as described above. Now there is an overlap between the 1-alleles on Y_1+Y_2-L = 3+5–6 = 2 loci:

$$\begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}0\\1\\1\\1\\1\\1\end{pmatrix}$$

L-Y_1 of the Y_2 maternal 1-alleles can be paired with paternal 0-alleles. The remaining Y_2-(L-Y_1) maternal 1-alleles must be paired with paternal 1-alleles.

The relations above give the minimum possible number of pairs given the union of gametes of classes Y_1 and Y_2. The next lowest number of pairs occurs if we exchange a 0 on the paternal gamete that is matched with a 1 on the maternal gamete with a 1 on the paternal gamete that is matched with a 0 on the maternal gamete. The resulting paternal gamete will be of the same gamete class, but there is now a new (0 0) pair and a new (1 1) pair in the resulting genotype. Every time we switch alleles like this we create a (0 0) and a (1 1) pair.

The pairing of gametes in classes Y_1 and Y_2 that produces the maximum number of pairs occurs when the maternal gamete has 1-alleles on loci 1 to Y_2 (with the 1-alleles on the Y_1 gamete still arranged as described above). Then the number of (1 1) pairs is equal to Y_1 and the number of (0 0) pairs is equal to L-Y_2 (recall that $Y_1 < Y_2$).

Example 3: Take $L = 6$, $Y_1 = 2$, and $Y_2 = 3$, as in example 1. Now assume that the specific gametes are as follows:

$$\begin{pmatrix}1\\1\\0\\0\\0\\0\end{pmatrix} \quad \text{and} \quad \begin{pmatrix}1\\1\\1\\0\\0\\0\end{pmatrix}$$

The number of (11) pairs is Y_1 = two and the number of (00) pairs is L-$Y_2 = 3$.

In summary, if $Y_1 + Y_2 \leq L$, then X_1 ranges from 0 to Y_1 and X_0 ranges from L-(Y_1+Y_2) to L-Y_2. If $Y_1 + Y_2 \geq L$, then X_1 ranges from (Y_1+Y_2)-L to Y_1 and X_0 ranges from 0 to L-Y_2. [A6]

Thus, for any given $[Y_1,Y_2]$ we have determined the set of possible values for X_0 and X_1.

The last thing that we need is an expression for $P([X_0,X_1]|[Y_1,Y_2])$ when $[X_0,X_1]$ is possible given $[Y_1,Y_2]$. This requires application of simple combinatorics:

$$P([X_0, X_1]|[Y_1, Y_2]) = \frac{\begin{pmatrix}\text{number of ways}\\ \text{to arrange the (11)}\\ \text{pairs on } L \text{ loci}\end{pmatrix}\begin{pmatrix}\text{number of ways to}\\ \text{arrange the (00) pairs}\\ \text{on the remaining loci}\end{pmatrix}\begin{pmatrix}\text{number of ways to}\\ \text{arrange the (01) pairs}\\ \text{on the remaining loci}\end{pmatrix}}{\begin{pmatrix}\text{The total possible number of ways}\\ \text{to make an } L - \text{ locus genotype out}\\ \text{of } Y_1 \text{ and } Y_2 \text{ parental gametes}\end{pmatrix}}$$

$$= \frac{\binom{L}{X_1}\binom{L-X_1}{X_0}\binom{L-X_1-X_0}{Y_1-X_1}}{\binom{L}{Y_1}\binom{L}{Y_2}} \quad [A7]$$

Note that there is another factor, number of ways to arrange the (0 1) pairs on the remaining loci, which equals one and is not therefore shown. By including this term and writing things out in terms of factorials we can show the symmetry of this expression with respect to Y1 and Y2. The numerator is the total number of ways to get an [X0,X1] genotype from parental gametes Y1 and Y2. The denominator is the total number of genotypes that can be made from parental gametes of classes Y1 and Y2.

Returning to equation A3, we have all of the pieces (equations A4, A5, A7 and the set of possible values of $X0$ and $X1$ given by A6), and the recursion for $P(Y,t)$ can programmed in a straightforward manner. This algorithm has been tested against the standard two and three locus models with selection.

Appendix 2. Example of an All-Male and Male-Female Releases

To illustrate the details of releases, we give an example of an all-male and a male-female release.

2:1 All-Male Release. Assume that the wild population has 50 males and 50 females. The release population has 200 males. The frequency of the no-conditional lethal genotype immediately after the 2:1 release is then 33% and the frequency of all-conditional lethal genotype is 67%. There is a total of 250 males, with 80% being released-type and 20% being wild type. The gametic disequilibrium of the no-conditional lethal (CL) gamete type in the release generation is

$$D = (freq.\ of\ no - CL\ gamete\ type) - (1 - CL\ gene\ freq)^L \quad [A8]$$

$$D = 0.33\text{-}0.33^L \quad [A9]$$

No Selection. If there is no selection against the released type insects, then the release generation matings will be 40 matings of type (wild female × released male) and 10 matings of type (wild female × wild male).

The offspring will then be 80% of genotype AaBbCc. . . . and 20% of genotype aabbcc. . . , where capital letters denote a conditional lethal allele and lowercase letters denote the absence of a conditional lethal allele. The frequency of the conditional lethal allele is $0.5 \times 0.8 = 0.4$. The gametic disequilibrium in the gametes going into the F_1 generation is

$$D = (\textit{freq. of no} - \textit{CL gamete type}) - (1 - \textit{CL gene freq})^L$$

$$D = (0.8*0.5 + 0.2) - 0.4^L = 0.6 - 0.6^L \quad [A10]$$

It is clear that A10 is greater than A9 for all L. This happens because all no-conditional lethal females mate and therefore the proportion of no-conditional lethal gametes that go into matings is high. From this, we can understand the increase in gametic disequilibrium in the F1 generations seen in Figs. 2b and 3d.

Selection Against Conditional Lethal Allele. If we define w_{CL} as the fitness of the all-conditional lethal genotype and w_{no} as the fitness of the no-conditional lethal genotype, then the matings in the release generation are as follows: (50) (0.8) (w_{CL}) matings of type (wild female × released male) and $50(1-0.8)w_{no}$ matings of type (wild female x wild male).

With no selection, the frequency of the no-conditional lethal genotype dropped from 0.33 to 0.20 between the release and F1 generations (see Fig. 2a). If w_{CL} is sufficiently small, then $(1-0.8)\ w_{no}$ will exceed 0.33, and we will see an increase in the no-conditional lethal genotype frequency in the F1 generation. This increase will be greater for larger L because w_{CL} decreases with increasing L (as reflected, for example, in Figs. 3a and 5a).

2:1 Male-Female Release. The male-female release is simpler. We again assume that the wild population consists of 50 males and 50 females. Then the released population has 100 males and 100 females. 33% of each gender is wild and 67% of each gender is released. The release generation matings are as follows: $(50)w_{no}(0.67)w_{CL}$ (wild female × released male), $(50)\ w_{no}\ (1-0.67)w_{no}$ (wild female × wild male), $(100)\ w_{CL}\ (0.67)\ w_{CL}$ (released female × released male), and $(100)\ w_{CL}\ (1-0.67)\ w_{no}$ (released female × wild male).

If there is no selection against the conditional lethal allele (thus $w_{no} = w_{CL} = 1$), then the fraction of no-conditional lethal genotypes will be $50(1-0.67)/300 = 0.05$. Recall that the all-male release had a no-conditional lethal genotype frequency of 0.2 in the F_1 generation with no selection. However, note that the number of matings producing the no-conditional lethal genotype is higher in the male-female release than the all-male release (16 versus 10). The frequency is lower because the total number of matings is higher when females are released. This explains why the no-conditional lethal frequency is higher in the F_1 generation in the all-male release in Fig. 8 than in the male-female release.

If there is selection against the conditional lethal allele, then the frequency of the no-conditional lethal genotype in the F_1 generation will be $(50/300)(w_{no})(1-0.67)$. This increases as increases (because w_{no} increases and w_{CL} decreases).

A Cluster of Vitellogenin Genes in the Mediterranean Fruit Fly *Ceratitis capitata*: Sequence and Structural Conservation in Dipteran Yolk Proteins and Their Genes

M. Rina* and C. Savakis*,†

**Institute of Molecular Biology and Biotechnology, Research Center of Crete, Foundation of Research and Technology, Heraklion, Crete, Greece and †Division of Medical Sciences, Medical School, University of Crete, Crete, Greece*

Manuscript received September 27, 1990
Accepted for publication November 29, 1990

ABSTRACT

Four genes encoding the major egg yolk polypeptides of the Mediterranean fruit fly *Ceratitis capitata*, vitellogenins 1 and 2 (VG1 and VG2), were cloned, characterized and partially sequenced. The genes are located on the same region of chromosome 5 and are organized in pairs, each encoding the two polypeptides on opposite DNA strands. Restriction and nucleotide sequence analysis indicate that the gene pairs have arisen from an ancestral pair by a relatively recent duplication event. The transcribed part is very similar to that of the *Drosophila melanogaster* yolk protein genes *Yp1*, *Yp2* and *Yp3*. The *Vg1* genes have two introns at the same positions as those in *D. melanogaster Yp3*; the *Vg2* genes have only one of the introns, as do *D. melanogaster Yp1* and *Yp2*. Comparison of the five polypeptide sequences shows extensive homology, with 27% of the residues being invariable. The sequence similarity of the processed proteins extends in two regions separated by a nonconserved region of varying size. Secondary structure predictions suggest a highly conserved secondary structure pattern in the two regions, which probably correspond to structural and functional domains. The carboxy-end domain of the *C. capitata* proteins shows the same sequence similarities with triacylglycerol lipases that have been reported previously for the *D. melanogaster* yolk proteins. Analysis of codon usage shows significant differences between *D. melanogaster* and *C. capitata* vitellogenins with the latter exhibiting a less biased representation of synonymous codons.

THE major egg yolk proteins (vitellogenins) of higher Diptera are polypeptides of 44,000 to 50,000 daltons and differ from those of other egg laying animals. In contrast, the vitellogenins from species as diverse as the locust, nematode, frog and chicken probably have a common evolutionary origin. They are generally larger in size, are encoded by multigene families and share amino acid sequence similarities; no significant sequence similarities can be detected between them and the dipteran yolk proteins (SPIETH *et al.* 1985; NARDELLI *et al.* 1987). The latter show local amino acid sequence similarity with members of the triacylglycerol lipase family (BOWNES *et al.* 1988).

The best studied dipteran vitellogenins are the three yolk proteins of *Drosophila melanogaster*. These polypeptides, designated YP1, YP2 and YP3, are synthesized in the fat body of adult females, secreted in the hemolymph and taken up by developing oocytes. In addition, *D. melanogaster* vitellogenins are synthesized in developing follicular epithelial cells and transported directly into the oocyte (BOWNES and HAMES, 1978; WARREN and MAHOWALD, 1979; BRENNAN *et al.* 1982). Their transcription is stimulated by ecdysteroid hormones; β-ecdysone induces yolk protein synthesis in the fat body of adult males, which normally do not produce yolk proteins (POSTLETHWAIT, BOWNES and JOWETT 1980). The three proteins are highly related to each other and are encoded by single-copy genes (*Yp1*, *Yp2* and *Yp3*) which are localized on the *X* chromosome. Genes *Yp1* and *Yp2* are transcribed divergently and are separated by 1.2 kb of DNA, while *Yp3* is located approximately one megabase away (HUNG and WENSINK 1983; GARABEDIAN *et al.* 1987; YAN, KUNERT and POSTLETHWAIT 1987). Tissue-specific transcriptional enhancers common for both genes are localized in the *Yp1/Yp2* intergenic region (GARABEDIAN, HUNG and WENSINK 1985; GARABEDIAN, SHEPHERD and WENSINK 1986). A very similar arrangement was described for the vitellogenin genes of *Drosophila grimshawi*, a Hawaiian endemic species which belongs to a different sub-genus than *D. melanogaster*. This species has three genes which cross-hybridize to each other and to the *melanogaster Yp* genes; S1 nuclease analysis has shown that two of the genes are closely linked and transcribed with opposite orientations with their 5′ ends 1.75 kb apart (HATZOPOULOS and KAMBYSELLIS 1987).

The *D. melanogaster* vitellogenins are characterized by three regions which show extensive primary and secondary structure homology among the genes but not between each other, separated by a non-conserved

Exhibit B

regions with variable length (HUNG and WENSINK 1983). One of the conserved regions has significant sequence similarity with part of the lipid-binding domain of triacylglycerol lipases; it has been suggested that this region has a lipid-binding function (BOWNES *et al.* 1988; PERSSON *et al.* 1989).

The vitellogenins of *Ceratitis capitata* have been studied to considerable detail. This species has two major yolk polypeptides, designated VG1 and VG2, with molecular weights of 49,000 and 46,000 daltons respectively. They have been purified and show immunological cross-reactivity with the *D. melanogaster* homologs (RINA and MINTZAS 1987, 1988). As in *D. melanogaster* (KOZMA and BOWNES 1986), they are synthesized in the fat body and follicular epithelial cells and are induced in males by β-ecdysone (RINA and MINTZAS 1988).

The Mediterranean fruit fly *C. capitata* (family Tephritidae; medfly) is a higher dipteran which presents several important advantages as an organism of choice for comparative molecular studies with *D. melanogaster*. It is phylogenetically close enough to *D. melanogaster* to allow cloning of Drosophila gene homologs by interspecific nucleic acid hybridization, but distant enough for comparisons to be meaningful. Furthermore, it has been adapted easily to inexpensive laboratory culture, it has a 24-day life cycle, and has well characterized polytene chromosomes (ZACHAROPOULOU 1990). Last but not least, medfly is an insect of economic importance amenable to biological control such as the sterile male technique. Cloning of two chorion protein genes and one actin gene from *C. capitata* was reported recently (KONSOLAKI *et al.* 1990; TOLIAS *et al.* 1990; HAYMER *et al.* 1990). Of these, the actin gene and one of the chorion genes (*Ccs36*) were cloned by heterologous hybridization to *D. melanogaster* probes, while chorion gene *Ccs38* was cloned by a differential screening procedure; gene *Ccs38* was subsequently shown to cross-hybridize with the *D. melanogaster s38* gene.

In an effort to obtain and analyze *C. capitata* promoters that are expressed in only one sex, we cloned the genes encoding vitellogenins. Here we report the structure of these genes and the results of DNA sequence analysis.

MATERIALS AND METHODS

Flies and materials: A *C. capitata* strain obtained from A. MINTZAS (Department of Biology, University of Patras, Greece) was used for all experiments. The strain was originally established in the laboratory by P. A. MOURIKIS (Benakeion Institute of Phytopathology, Athens, Greece) with flies from the Southern Peloponnese (Greece) and Palermo (Italy). Insects were raised at 22–25° as described previously (MINTZAS *et al.* 1983). Adults were maintained on a one part sucrose to one part dried yeast diet. Under these conditions, embryonic development lasts about 48 hours, and the complete life cycle of the insect is approximately 24 days.

^{32}P-Labeled nucleotides were from Amersham, and restriction and modification enzymes from MinoTech (Heraklion), Pharmacia and Bethesda Research Laboratories.

Construction and screening of the genomic library: DNA (20 μg) from 24-hr-old medfly embryos was partially digested with restriction endonuclease *Mbo*I, extracted with phenol/chloroform, precipitated with ethanol, and fractionated on a 10–40% sucrose gradient. Fractions containing fragments of 15–20 kb in length were retained for ligation to vector DNA. For ligation, 0.25 μg genomic DNA fragments were combined with 0.83 μg lambda EMBL4 arms produced by digestion of phage DNA with *Bam*HI. *In vitro* packaging was as described previously (MANIATIS, FRITSCH and SAMBROOK 1982). Approximately 250,000 plaques were screened, as described by BENTON and DAVIS (1977), by hybridization at 55° to a *D. grimshawi* cDNA clone (HATZOPOULOS and KAMBYSELLIS 1987) corresponding to the vitellogenin 1 mRNA of this species. This clone was used because it was available to us and had been shown to cross-hybridize strongly with its *D. melanogaster* homolog.

General methods: Genomic DNA was prepared essentially as described previously (HOLMES and BONNER 1973). Preparation of phage and plasmid DNA, agarose gel electrophoresis of DNA, and blotting to nitrocellulose membranes were carried out using standard procedures (MANIATIS, FRITSCH and SAMBROOK 1982). DNA probes were prepared by nick-translation (MANIATIS, FRITSCH and SAMBROOK 1982) or by random hexanucleotide priming (FEINBERG and VOGELSTEIN 1983). Hybridizations of ^{32}P-labeled probes to blotted nucleic acids were performed as described by MANIATIS, FRITSCH and SAMBROOK (1982), at 38° (*D. grimshawi* probe) or 42° (*C. capitata* probes) in 50% formamide, 5× SSC, 0.5% SDS, 10 mM EDTA, 100 μg/ml sonicated, heat-denatured herring sperm DNA, and 5× DENHARDT (1966) solution. DNA sequencing was done by the double stranded dideoxy chain termination method (WALLACE *et al.* 1981).

DNA and protein sequence analysis: The program packages ANALYSEQ and ANALYSEP (STADEN 1984) were used for sequence analysis. Optimal alignment of protein sequences was carried out by the IALIGN program (DAYHOFF, BARKER and HUNT 1983) of the Protein Identification Resource (National Biomedical Research Foundation) and by the multiple alignment program CLUSTAL (HIGGINS and SHARP 1988). All programs were run on a VAX/VMS computer. Graphics were processed and plotted with a Macintosh microcomputer running a terminal emulation software.

RESULTS AND DISCUSSION

Cloning of *C. capitata* vitellogenin genes: Figure 1 shows the results of a hybridization experiment in which vitellogenin DNA from another higher dipteran, *D. grimshawi*, was used to probe a *C. capitata* genomic DNA blot at low hybridization stringency. The probe was a cDNA clone (plasmid clone c357) which contains approximately 900 nt. corresponding to the carboxy terminus half of the Vitellogenin 1 mRNA from *D. grimshawi* (HATZOPOULOS and KAMBYSELLIS 1987). At least four prominent *Hin*dIII fragments and two *Eco*RI fragments were detected by this probe in *C. capitata* DNA. The same probe hybridized

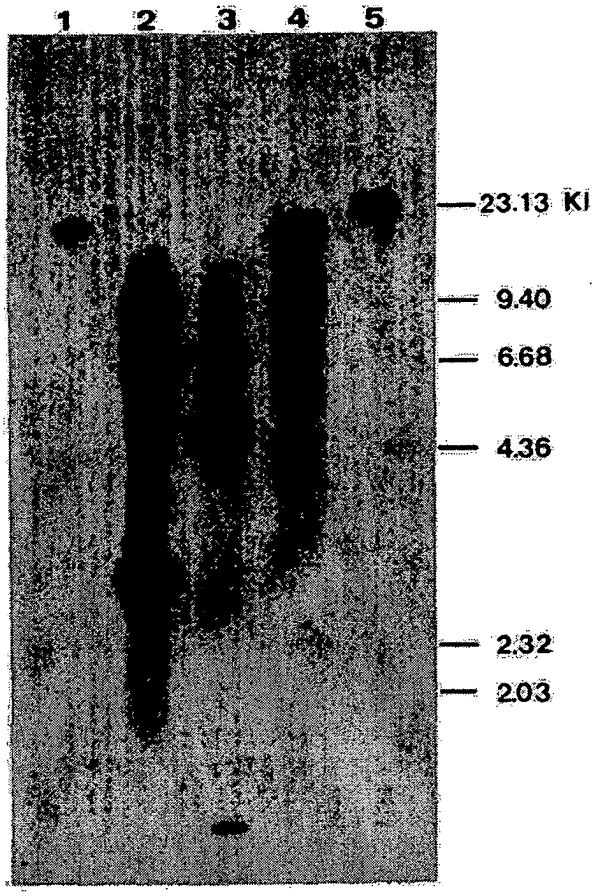


FIGURE 1.—Southern blot hybridization of *D. grimshawi* vitellogenin DNA to *C. capitata* genomic DNA. Embryonic DNAs from *D. grimshawi* (lane 2) and *C. capitata* (lanes 3 and 4) were digested with restriction enzymes, separated on a 0.9% agarose gel and blotted on a nitrocellulose membrane filter. The blot was hybridized at low stringency (5× SSC, 38°, 50% formamide) with nick-translated plasmid c357 DNA. Lanes 2 and 3, DNA digested with *Hin*dIII. Lane 3, DNA digested with *Eco*RI. Lanes 1 and 5 are size standards.

to three *Hin*dIII fragments of *grimshawi* DNA, each corresponding to one of the three vitellogenin genes identified in this species (HATZOPOULOS and KAMBYSELLIS 1987).

To clone the genomic vitellogenin-related sequences, a *C. capitata* DNA library was constructed in vector EMBL4 using genomic DNA fragments produced by partial digestion with restriction endonuclease *Mbo*I. Approximately 2.5×10^5 recombinant phage plaques were screened for hybridization to the *D. grimshawi* vitellogenin cDNA probe. Five clones gave strong signal and were isolated and characterized by a combination of restriction mapping and blot hybridization. Figure 2 is a summary of the results of this analysis.

The restriction maps of three of these clones (ccv71, ccv51 and ccv53) were clearly overlapping. The short overlap between clone ccv72 and clones ccv51 and ccv53 was confirmed by Southern blot hybridization analysis (data not shown). The four overlapping clones cover a 37 kilobase region of the *C. capitata* genome which contains four distinct small regions of vitellogenin DNA-related sequences. Southern analysis of these clones showed four non-contiguous fragments hybridizing strongly to the *D. grimshawi* cDNA probe: a 0.56-kb *Hin*dIII/*Bam*HI and a 1.4-kb *Hin*dIII/*Hin*dIII fragment from clone ccv72, and a 0.93 *Hin*dIII/*Bam*HI and a 0.56-kb *Hin*dIII/*Bam*HI fragment from the other three clones. The four regions corresponding to these fragments were designated α, β, γ and δ (Figure 2).

Two lines of evidence strongly suggest that the restriction fragments hybridizing to the *D. grimshawi* vitellogenin probe contain medfly vitellogenin gene sequences. First, hybridization of one of these fragments (the 0.93 *Hin*dIII/*Bam*HI fragment from clone ccV71 corresponding to region γ) to total fat body RNA from staged female and male medflies gave a single band in 36-hour or older females, but not in males (RINA and MINTZAS 1988); this pattern accurately reflects the levels of translatable vitellogenin mRNA in fat body of the developing medfly. Second, hybrid-selected translations showed that mRNAs hybridizing to clone ccV71 DNA are translated into two polypeptides with the same mobilities in SDS-polyacrylamide gels as newly synthesized vitellogenins 1 and 2 (data not shown).

The arrangement of the vitellogenin-related regions α to δ, combined with the observed symmetry of some of the restriction sites around these regions, suggests the presence of a cluster of duplicated genes. This was confirmed by sequence analysis (see below). The restriction map of the fifth clone, ccv81, is clearly not overlapping with that of the 37-kb region covered by the other clones. However, part of the ccv81 map is indistinguishable from that of the 37-kb region between positions at 20 and 31 kb (Figure 2; clone ccv81 has been aligned accordingly). This clone may represent an adjacent duplication of the γ-δ gene pair, may correspond to a third pair of vitellogenin genes, not directly linked to the α-δ cluster, or, alternatively, could be the result of a cloning artefact. It is notable that in genomic DNA blots, four *Eco*RI fragments (approximately 6.0, 9.5, 15.0 and >20.0 kb in size) are detected with vitellogenin DNA probes (Figure 1). Since only the 15- and >20-kb fragments could be accounted for by the composite map of the α-δ cluster and by the restriction map of clone ccv81, it is possible that more vitellogenin genes exist in the medfly genome. However, all cloned medfly vitellogenin genes are probably closely linked, since clones ccv72, ccv71 and ccv81 hybridize *in situ* to the same band on chromosome 5 of the medfly (A. ZACHAROPOULOU, M. FRISARDI, A. ROBINSON and C. SAVAKIS, manu-

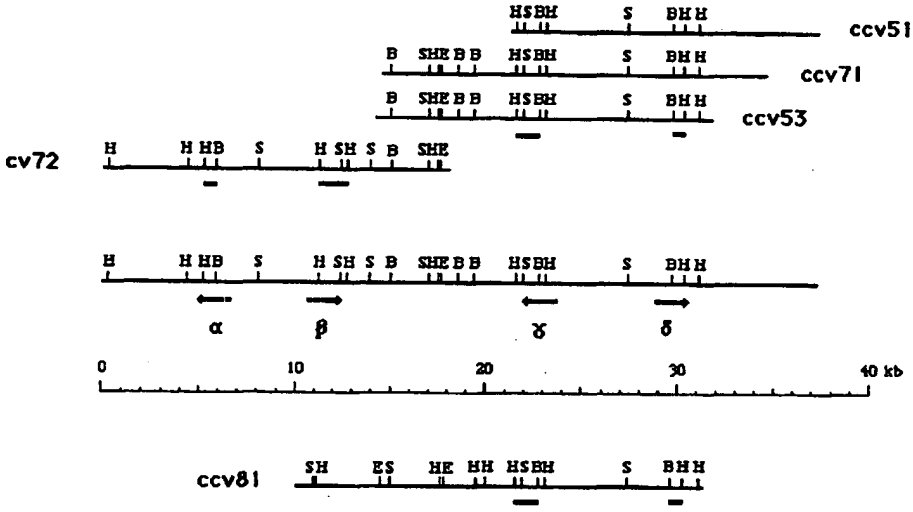


FIGURE 2.—Restriction maps and arrangement of *C. capitata* genomic clones containing vitellogenin-related sequences. Restriction maps of the four overlapping clones ccv72, ccv53, ccv51 and ccv71 are shown at the top. The sites shown are: *Bam*HI (B), *Sal*I (S), *Hin*dIII (H) and *Eco*RI (E). The bars below clones ccv72 and ccv53 correspond to the fragments produced by double *Bam*HI/*Hin*dIII digestion that hybridize to the *D. grimshawi* probe. The composite map of the cluster is shown below the clones. The four vitellogenin genes (α to δ) were identified by sequencing (see text). The restriction map of the non-overlapping clone ccv81 is shown below the size scale.

script in preparation). We have concentrated our analysis on the genes of the α-δ cluster.

Structure and chromosomal localization of the γ and δ genes; conserved features between *C. capitata* and *D. melanogaster*: Sequencing of the γ and δ regions showed that each contains a gene highly homologous to the *D. melanogaster* yolk protein genes. Figure 3 shows the sequence of 2364 bp of genomic DNA covering the γ region. The sequence extends from base no. +953 to base −1411 relative to the *Hin*dIII site at 23 kb of the composite map shown in Figure 2. Conceptual translation in all six frames showed three open reading frames (all in the orientation indicated in Figure 2) with significant similarity to the *D. melanogaster* yolk polypeptides, suggesting the presence of two introns. By aligning the derived polypeptide sequences to the available *D. melanogaster* yolk protein sequences we arrived at the intron/exon structure indicated in Figure 3. The first coding part begins with an ATG at base 433 of the sequence and ends at position 655, which is the first base of codon 74. The nucleotide sequence surrounding the initiation codon (C A A C A T G) is in good agreement with the consensus sequence C/A A A A/C A T G flanking translational start sites in Drosophila (CAVENER 1987). A 67-bp intron separates the first coding part from a second, 389-bp exon beginning at base 723. A second intron is placed between bases 1112 and 1179. The 5′ and 3′ ends of both introns conform to consensus sequences (G T A/G A G T ... Y N Y Y Y Y N Y A G) (MOUNT 1982; TEEM *et al.* 1984); in addition, both introns contain versions of the internal splice signal C/T T A/G A C/T (KELLER and NOON 1984) upstream from the 3′ splice site. The third coding part is 699 bp long, ending with a TAA at base 1879. Two tandem repeats of the consensus polyadenylation signal sequence A A-T A A A A (PROUDFOOT and BROWNLEE 1976) are located 100 nucleotides downstream from the termination codon.

The 1702-bp sequence from the δ region is shown in Figure 4. This sequence extends from base −1137 to base +565 relative to the *Bam*HI site at approximately 30.3 kb of the composite map shown in Figure 2. The sequence contains two open reading frames which are read in the opposite direction from the δ gene and also show considerable sequence similarity to the *D. melanogaster Yp* genes. The proposed intron/exon structure of the gene is as follows. The first coding part is 211 bp long and begins with the ATG at base 237 of the sequence. This ATG is also embedded within a sequence similar to the Drosophila consensus (see Figure 6A). The exon ends at position 447, which is the first base of codon 71, and is followed by a 89-bp intron with acceptable splice signals. The second coding part is 1055 bp long, ending with a TAA codon at base 1592.

Although the proposed intron/exon structures of the γ and δ genes were not subjected to direct testing, such as S1 protection experiments or cDNA sequence analysis, we believe that they represent the correct structures because of their striking similarity to the *D. melanogaster* yolk protein genes. *D. melanogaster* has three genes, *Yp1*, *Yp2* and *Yp3*, coding for the three yolk proteins found in this species. The structures and the arrangement of consensus sequences are very similar in these genes. *Yp1* and *Yp2* have a short exon followed by a single short intron (75 bp in *Yp1* and 68 bp in *Yp2*) and then by a longer exon. *Yp3* has two short introns of 62 and 72 bp; the first is at the same position as the intron in *Yp1* and *Yp2*. The 5′ consensus sequences of the three genes (TATA box, capping site and translation initiation consensus sequence) are also at very similar positions; in addition, all three genes have rather short (51 to 61 bp) 5′ untranslated regions (HUNG and WENSINK 1983; GARABEDIAN *et*

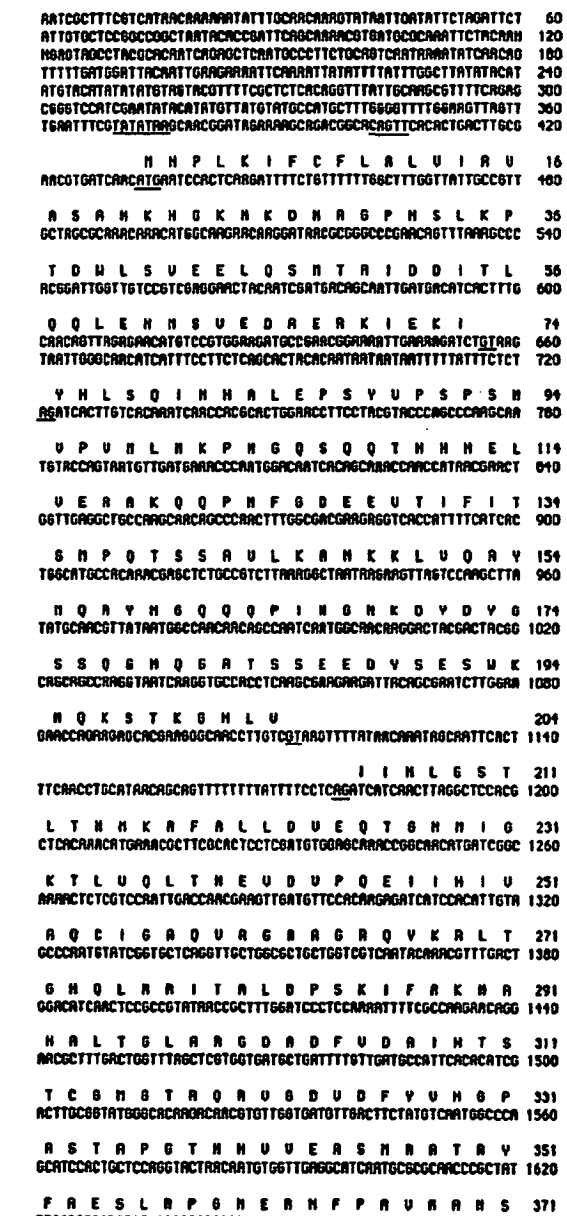


FIGURE 3.—Nucleotide sequence of the *Vg1-γ* gene and surrounds. The predicted amino acid sequence is shown above the DNA sequence. The putative TATA box, cap site, initiation and termination codons, polyadenylation signals, and first and last two bases of each intron are underlined. Numbers refer to nucleotide and amino acid positions.

al. 1987; YAN, KUNERT and POSTLETHWAIT 1987). The two sequenced medfly genes have the following structural features in common with their *D. melanogaster* homologues:

Sequence alignment and position of introns: Removal of the two introns from the γ gene gives an open reading frame of 1311 nucleotides, which can encode a 437 amino acid polypeptide with a molecular weight of 48,122 daltons. Respectively, the δ gene can encode a 422 amino acid polypeptide with a molecular weight of 45,434 daltons. Medfly vitellogenins 1 and 2 have molecular weights of 49,000 and 46,000 daltons, respectively (RINA and MINTZAS 1987). Based on the good agreement between the calculated and the observed molecular weights, we suggest that the γ gene encodes VG1 and the δ gene encodes VG2. Figure 5 is an alignment of these deduced polypeptide sequences to the known sequences of the three *D. melanogaster* yolk proteins. The five sequences can be aligned with only two major gaps, which are located within exons. This alignment illustrates that the position of the introns is strictly conserved in the five genes: Intron 1 of the *Vg1* gene and the single intron of the *Vg2* gene are at the same position as the intron present in all *D. melanogaster* yolk protein genes, *i.e.*, after the first base of a codon for tyrosine (HUNG and WENSINK 1983); intron 2 of the *Vg1* gene is at the same position as the second intron of the *D. melanogaster Yp3* gene (GARABEDIAN *et al.* 1987). With respect to intron/exon structure, therefore, the *Vg1* gene appears to be homologous to the *Yp3* gene, while the *Vg2* gene appears to be homologous to the *Yp1* and *Yp2* genes.

Size of introns: The two introns in the *Vg1* gene and the single intron in the *Vg2* gene are, as in the *D. melanogaster* yolk protein genes (HUNG and WENSINK 1983; GARABEDIAN *et al.* 1987), very short: 67 and 89 nucleotides for intron 1 of the *Vg1* and *Vg2* genes respectively, and 68 nucleotides for intron 2 of *Vg1*. Figure 6B shows a comparison of intron sizes and consensus splicing sequences between the vitellogenin genes of *C. capitata* and *D. melanogaster*.

5' Consensus sequences: The *D. melanogaster* yolk protein genes have, as most eukaryotic genes, a TATA (Hogness-Goldberg) box beginning at position −29 to −32 relative to the capping nucleotide (HUNG and WENSINK 1983; GARABEDIAN *et al.* 1987; YAN, KUNERT and POSTLETHWAIT 1987). In addition, the capping sites of all three genes match the insect cap site consensus sequence (HULTMARK, KLEMENTZ and GEHRING 1986). The γ gene has the sequence T A T A T A A between bases −61 and −55 relative to the initiation ATG; the δ gene has a T A T A A A A

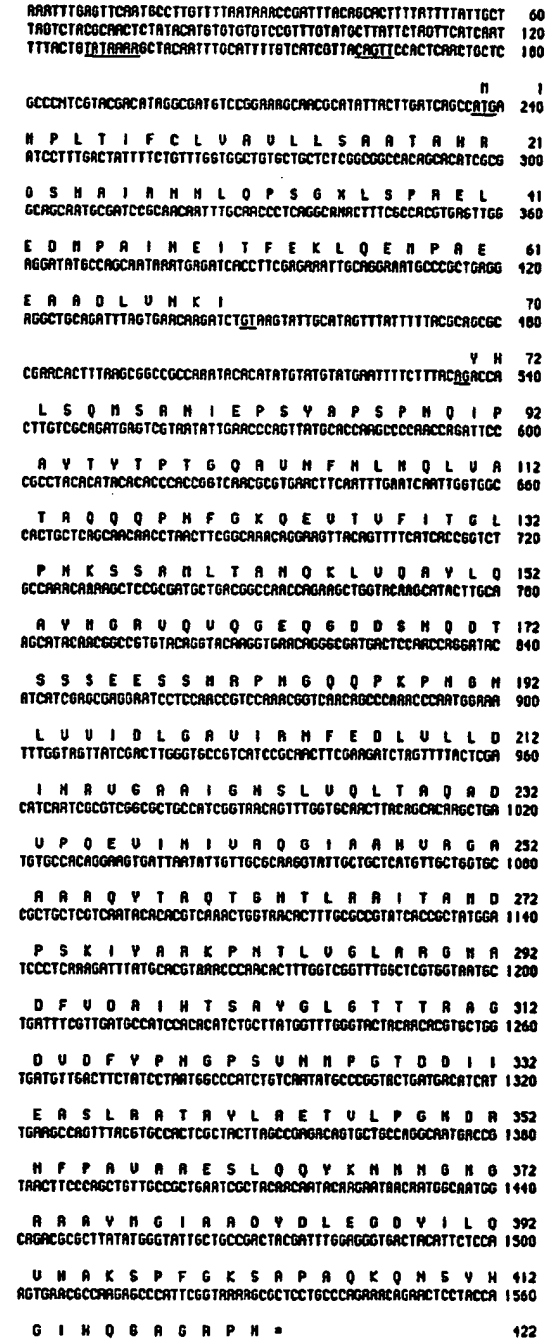


FIGURE 4.—Nucleotide sequence of the *Vg2-δ* gene and surrounds. The predicted amino acid sequence is shown above the DNA sequence. The putative TATA box, cap site, initiation and termination codons, polyadenylation signals, and first and last two bases of the intron are underlined. Numbers refer to nucleotide and amino acid positions.

between −109 and −103 from the ATG. Alignment of the five nucleotide sequences at the TATA boxes showed that each of the medfly genes has a capping site-like sequence at the canonical distance downstream of its putative TATA box. The γ gene has the sequence C A C A G T T and the δ gene has the sequence T A C A G T T 30 and 32 bases downstream from the TATA, respectively. These heptanucleotides represent 5/7 matches to the consensus insect cap site A T C A G/T T C/T. These features are shown in Figure 6A.

Size of the 5′ untranslated regions: The *D. melanogaster* yolk protein mRNAs have rather short 5′ leaders, 61, 51 and 56 nucleotides for *Yp1*, *Yp2* and *Yp3*, respectively (HUNG and WENSINK 1983; GARABEDIAN *et al.* 1987; YAN, KUNERT and POSTLETHWAIT 1987). If the putative capping sites identified are used by the γ and δ genes, then their 5′ sequences are also rather short, *ca.* 30 and 76 nucleotides, respectively.

Similarities in the 5′ flanking sequences: As in *D. melanogaster*, the sequence homology observed in the coding parts of the *C. capitata* vitellogenin genes does not extend into the 5′ flanking DNA. However, several short nucleotide sequences have been identified in *D. melanogaster*, which are repeated several times in the 5′ flanking DNA of the yolk protein genes (GARABEDIAN *et al.* 1987; YAN, KUNERT and POSTLETHWAIT 1987). The heptamer A/T A/T T G C A A or its complement is encountered seven times within 800 bp upstream from the *Yp3* gene, and five times in the 1.2 kb intergenic region between *Yp1* and *Yp2* (YAN, KUNERT and POSTLETHWAIT 1987). Matches to this heptamer or its complement are found at positions −88 and −343 (relative to the first T of the TATA box) of the γ gene, and at positions −62 and −72 of the δ gene (Figure 6C). A comparison between *D. melanogaster* and *C. capitata* flanking sequences also showed the presence of single copies of the sequence G A G N T C A A G/T G/T C G/C at distances from −575 to −124 relative to the TATA box in the *Yp2*, *Yp3*, γ and δ genes (Figure 6D). The consensus, and indeed three of the four actual sequences, are close relatives (9 of 12 nucleotides) to the sequence G G G T T C A A T G C A found at the ecdysone responsive element of the *D. melanogaster hsp27* gene promoter (RIDDIHOUGH and PELHAM 1987). Although the significance of these similarities is not known, transformation experiments in which *in vitro* modified medfly genes are introduced into the *D. melanogaster* genome would test the possibility that they represent regulatory elements conserved between *D. melanogaster* and *C. capitata*.

Another similarity between Drosophila and medfly vitellogenin genes is their position in the genome. In Drosophila these genes are on the *X* chromosome, while in medfly they are located on chromosome *5*. The *X* chromosome of the medfly is heterochromatic (ZACHAROPOULOU 1990), and recent *in situ* hybridization studies of medfly polytene chromosomes have revealed that several medfly genes homologous to

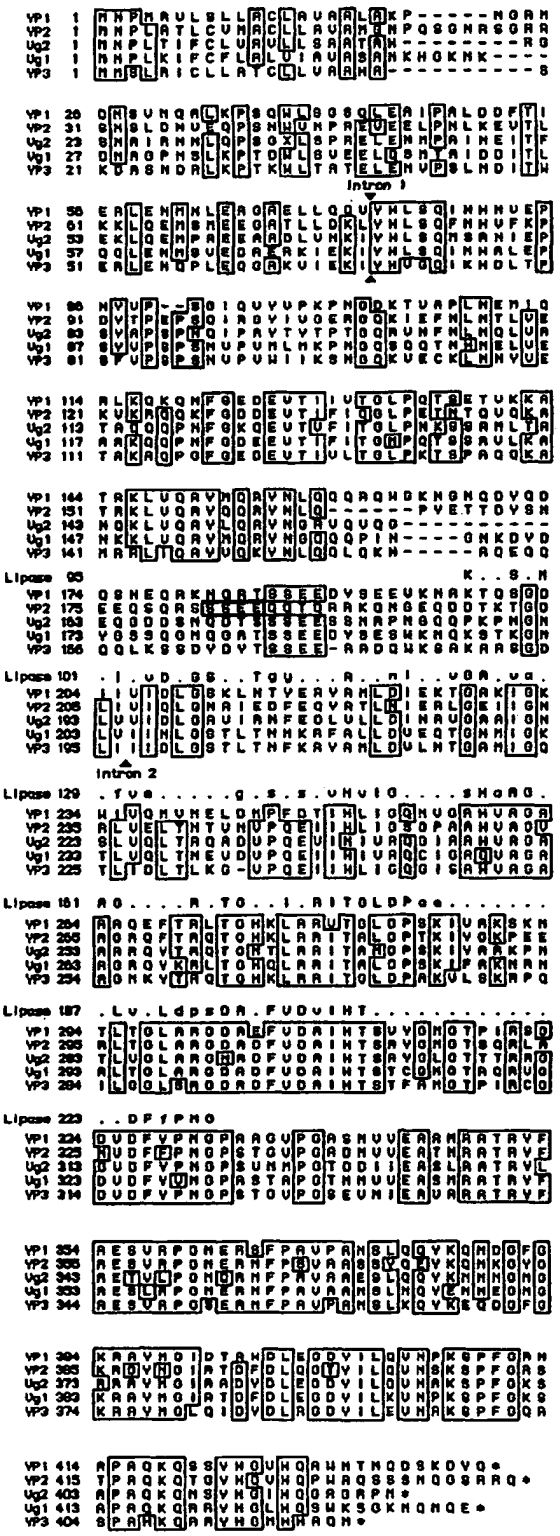


FIGURE 5.—Comparison of the amino acid sequences of the three *D. melanogaster* yolk polypeptides to the deduced *C. capitata* vitellogenin polypeptide sequences. YP1, *D. melanogaster* yolk protein 1 (NBRF database accession No A03332); YP2, *D. melanogaster* yolk protein 2 (NBRF database accession No A03333); YP3, *D. melanogaster* yolk protein 3 (sequence from GARABEDIAN *et al.* 1987); VG1 and VG2: *C. capitata* vitellogenin 1 and 2, (genes γ and δ) respectively. The five proteins were aligned using the program CLUSTAL (HIGGINS and SHARP 1988). The positions of the gaps were first determined by two-way comparisons using program IALIGN (DAYHOFF, BARKER and HUNT 1983) and then adjusted manually for maximum similarity. Residues which are identical in at least four of the sequences are boxed. The position of the introns is indicated by filled triangles (intron 2 is found only in *D. melanogaster Yp3* and in *C. capitata Vg1* genes). The similarity of these sequences to the pig triacylglycerol lipase (NBRF database accession No. A00732) is shown above the YP1 sequence. Capital letters show identities between lipase and at least two of the vitellogenins; lower-case letters indicate conservative replacements. Numbers indicate amino acid positions. Three small gaps in the lipase/vitellogenin alignment are not shown.

Drosophila *X*-linked genes (including the vitellogenins and the two chorion genes *s36* and *s38*) are located on chromosome *5* (A. ZACHAROPOULOU, personal communication). These syntenic associations, combined with those discovered for other medfly genes by MALACRIDA *et al.* (1986), further support Muller's hypothesis about the evolution of the Diptera (MULLER 1940) and add to the significance of comparisons between medfly and Drosophila.

The α and β genes have arisen from a recent duplication of the γ-δ pair: The restriction map of the α-β region shown in Figure 2 suggests that this region represents an inverted duplication of the γ-δ region. This was confirmed by partially sequencing the α and β loci.

Two parts of the α locus (429 and 565 bp) were sequenced. The first part is identical to bases 1 to 429 of the δ gene, with three differences: a 202A → G substitution at the 5′ untranslated region, a 260T → C conservative substitution at codon 8, and 363G → A, which results in a replacement, 43Asn → Ile. The second part is identical to bases 1138 to 1702 of the γ sequence, with three differences: A 1330T → C conservative substitution at codon 335, and two A → C changes at bases 1593 and 1594; these changes replace the TAA termination codon with a codon for serine, which is then followed by codons GAC, AAC and a termination codon, TAA. As a result, the carboxy terminus of the polypeptide coded by the α gene differs from that of the δ gene product by an extra SerAspAsn tripeptide.

Similar results were obtained by partially sequencing the β locus. The sequenced part (992 bp) is more than 98% identical to nucleotides 1130 to 2121 of the γ gene. Of the thirteen differences found, five are in intron 2 (1134A → T, 1153A → G, 1157G → A, 1168T → C, 1174C → G), seven are located downstream from the termination codon (1971T → G, 2056A → G, 2069C → G, 2071T → C, 2074T → C, 2093C → T, 2110G → A) and only one occurs in exon 3 (1187A → T, resulting in a replacement, 206Asn → Ile).

We conclude that the α-β and γ-δ pairs of genes

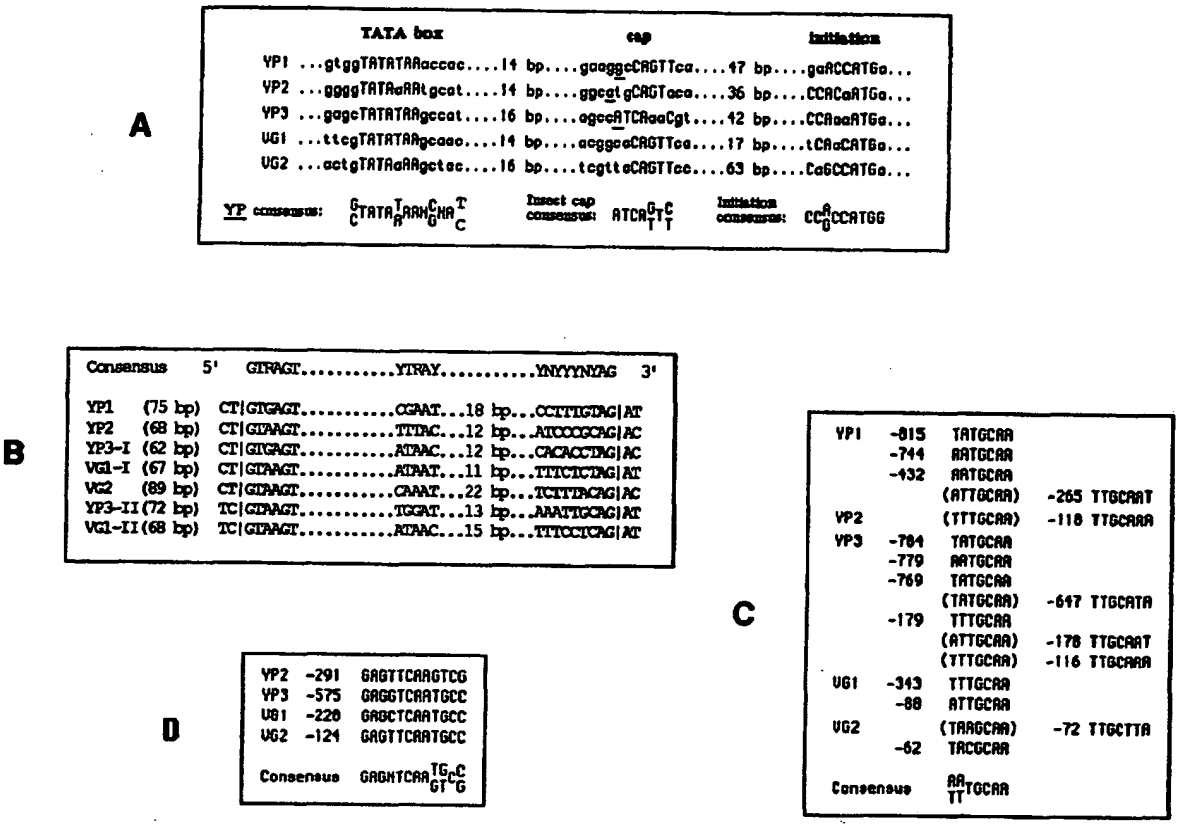


FIGURE 6.—Sequence similarities at noncoding regions in the vitellogenin genes of *D. melanogaster* and *C. capitata*. A, regions of nucleotide sequence similarity around the transcription and translation initiation sites. B, Comparison of the splice sites and intron sizes. C, Occurrences of the repeated heptamer (see text). D, Matches to the *D. melanogaster hsp27* ecdysone response element. For comparison, nucleotide positions in C and D are relative to the first T of the TATA or putative TATA boxes.

have arisen from a relatively recent duplication event. The data are not sufficient to rule out the possibility that α and β are pseudogenes. The findings, however, that the potentially coding regions have diverged less than the noncoding regions, and that all parts of the partial open reading frames which have been sequenced in α and β (450 codons in total) are free of stop codons, strongly suggest that these genes code for variants of the vitellogenins 2 and 1 respectively. This is also supported by sequence analysis of vitellogenin cDNA clones from medfly ovaries (K. PALIAKASIS and C. SAVAKIS, unpublished results).

***D. melanogaster* and *C. capitata* vitellogenin proteins show extensive sequence and structural conservation:** A striking degree of conservation between the *D. melanogaster* and the *C. capitata* vitellogenins is revealed when the five amino acid sequences are aligned for maximum similarity. This conservation pertains to primary sequence, hydrophobicity patterns and predicted secondary structure (Figures 5 and 7). For comparison, we divide each sequence into five regions (**a** to **e** in Figure 7):

The conserved amino terminal region of all the proteins (region **a**) is 19 or 20 residues long and hydrophobic (Fig. 7, bottom); we conclude this is the signal sequence for secretion (BLOBEL and DOBBERSTEIN 1975).

Region **b**, corresponding to residues 26 to 159 of YP1, is characterized by a low degree of sequence conservation: Twenty six residues (19%) are invariant in all five proteins. However, there are virtually no gaps in the alignment (only a two-residue deletion in YP1) and there are several conservative replacements, which result in a pronounced conservation of secondary structure and hydrophobicity patterns (Figure 7). A short region between regions **a** and **b** cannot be aligned without the introduction of insertions/deletions.

Region **c** corresponds to amino acids 160 to 201 of YP1, and shows no apparent conservation, with the exception of a SerSerGluGlu sequence shared by all proteins. This region varies in size, from 42 amino acids in YP1 to 33 amino acids in VG2. It contains many amino acids with charged and polar side chains (Figure 7, bottom), but does not seem to be conserved at the secondary structure level.

Region **d** is the most conserved one. It is 228 amino acids long, spanning residues 202 to 427 of YP1. The

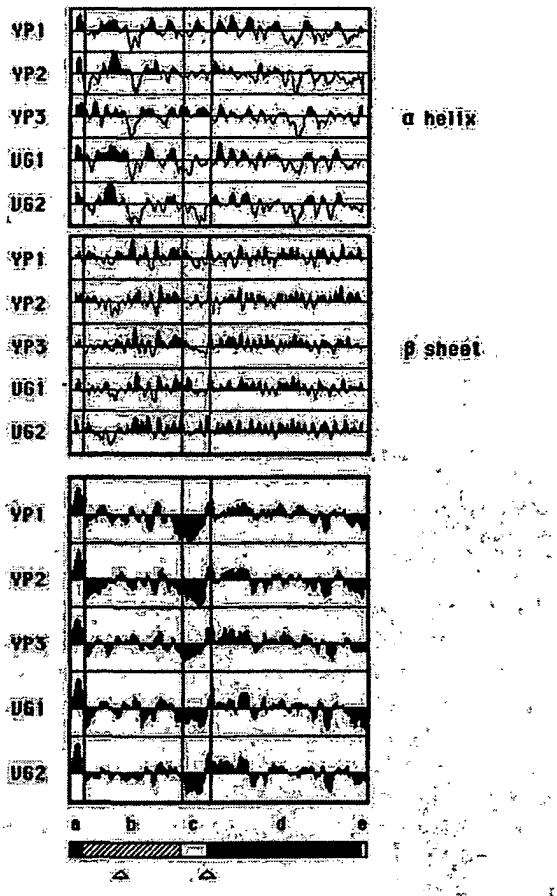


FIGURE 7.—Hydrophobicity and secondary structure predictions for the vitellogenins of *D. melanogaster* and *C. capitata*. (Top) Prediction of α-helix and β-sheet structure according to GARNIER *et al.* (1978). Regions predicted to have a given secondary structure are shown above the line and are shaded. (Bottom) Hydrophobicity plots of the five proteins. Regions above the line are hydrophobic, and those below are hydrophilic. Calculation and graphic representation were done using the ANALYSEP software package (STADEN 1984). Bars below the graphs represent the consensus vitellogenin sequence, with similarity regions a and e indicated. Open triangles indicate the positions of introns.

five sequences are aligned along this region with only a single amino acid deletion in the YP3 sequence. Ninety two positions (40%) are invariant, and 90 positions are occupied by structurally conservative residues. As a result, this region is predicted to have a highly conserved secondary structure and shows a highly conserved hydrophobicity pattern (Fig. 7). In addition, the amino-end two-thirds of region d shows sequence similarity with pig triacylglycerol lipase (Figure 5), as previously shown for the *D. melanogaster* yolk proteins (BOWNES *et al.* 1988).

Region e, corresponding to the carboxy end of the proteins, is not conserved and varies in size from 3 (YP3) to 14 amino acids (YP2).

The conservation of secondary structure previously observed between YP1 and YP2 has led to the proposition that secondary structure is important in the common functions of the two proteins (HUNG and WENSINK 1983). Our results favor this idea. The extensive conservation between species estimated to have diverged 120 myr ago (BEVERLY and WILSON 1984), means that these genes are under strong selective pressure. Although the vitellogenins are often envisaged as nutritional storage proteins, they have a number of properties which may put severe constraints on changes of their secondary, as well as tertiary structure. First, they are known to oligomerize. Second, they probably interact with specific receptors for active transport across the follicular epithelium. Such protein-protein interactions could have strict structural requirements. Finally, there is evidence that they may play a regulatory role in embryogenesis by binding conjugated ecdysteroid hormones (BOWNES *et al.* 1988).

The highly conserved region d is of particular interest. It is almost coincident with the large exon of YP3 and VG1, and is separated from the rest of the molecule by a stretch of hydrophilic amino acids with no apparent secondary structure. It has previously been shown that this region of the *D. melanogaster* proteins shows weak, but significant sequence similarity with the "central homology region" shared by all members of the triacylglycerol lipase family (BOWNES *et al.* 1988; Persson *et al.* 1989). Our results confirm and extend this observation; 25 out of 145 residues in this region are invariable between pig triacylglycerol lipase and all five vitellogenin sequences, while 65 residues show conservative amino acid replacements (Figure 5). The lipase "central homology region" lies within the lipid binding domain of the lipases. Because no catalytic activities have been attributed to the vitellogenins, it has been proposed that this region constitutes a lipid binding domain which serves a lipid carrier function; this prediction is supported by the finding that fatty acid-conjugated ecdysteroids are bound to purified *D. melanogaster* vitellogenins (BOWNES *et al.* 1988).

Our results, combined with those from *D. melanogaster* and *D. grimshawi*, indicate that the vitellogenins of higher diptera are highly adapted proteins which are subject to strong selective pressure. It is attractive to hypothesize that the conserved regions of b and d correspond to discreet structural domains, each derived from different ancestral peptides and being involved in a different function (or functions). In searches of the NBRF and SWISPROT databases with sequences corresponding to regions b and c we did not obtain any significant similarities with other known proteins. Region d, which is in part similar to the vertebrate lipases, is probably a domain involved, at least in part, in lipid binding. The function(s) of the other regions are unknown, and difficult to reveal, mainly because of the lack of point mutations in the

D. melanogaster vitellogenins. This drawback, which is probably caused by the presence of multiple vitellogenin genes in the genome, could be overcome by introducing *in vitro* mutagenized gene copies into the germline and studying their effects on the wild-type genes, as originally suggested by HERSKOWITZ (1987).

From an evolutionary point, it is remarkable that the vitellogenins of higher diptera have a different ancestry from all other known vitellogenins, such as those of the chicken, frog, locust, and nematode. The latter are generally longer proteins encoded by genes which are interrupted by many introns and share sequence similarities which indicate a common evolutionary origin. No sequence similarity can be detected between the dipteran and the other vitellogenins. It appears, therefore, that during the evolution of higher diptera the functions of the major yolk proteins were taken over by a gene which has common origin with the present day triacylglycerol lipases. It is intriguing that another protein of diptera, the enzyme alcohol dehydrogenase (ADH), has an analogous evolutionary history. The *D. melanogaster* enzyme is related to prokaryotic ribitol dehydrogenases and shows no sequence or intron/exon structure similarity to any of the sequenced eukaryotic ADHs. In contrast, the ADH proteins from yeast, plants and mammals are all related to each other, having presumably evolved from a common ancestral gene (JORNVALL, PERSSON and JEFFREY 1981; JORNVALL *et al.* 1984; SULLIVAN, ATKINSON and STARMER 1990). With the increasing accumulation of new protein sequences it will be interesting to see whether such events have occurred also during the evolution of other taxa.

***C. capitata* and *D. melanogaster* genes have different codon usage patterns:** Synonymous codons are not used with equal frequency and, often, genes from one species share similarities in codon usage (GRANTHAM *et al.* 1980, 1981). Moreover, in species showing non-random codon usage, individual genes differ from each other in the degree, rather than in the direction of the bias (SHARP *et al.* 1988). Nonrandom usage of alternative codons can be generated by biases in mutation patterns and by selection operating at the level of translation. Generally, selection for more efficient translation is probably driving the codon usage patterns in several genes of prokaryotes and yeast; in these species abundantly expressed genes show strong codon usage biases while weakly expressed genes have a more even synonymous codon representation (GOUY and GAUTIER 1982; IKEMURA 1985; SHARP and LI 1986). In mammals codon usage varies among genes mainly in (G + C) content, and specifically in the frequency of the dinucleotide CpG, which correlates with base composition around the gene and in introns (AORTA and IKEMURA 1988). *D. melanogaster* genes also show considerable codon usage bias (BODMER and ASHBURNER 1984; ASHBURNER, BODMER and LEMEUNIER 1984). On average, there is a strong deficiency of A and a weaker deficiency of T in the third position, and a more marked under-representation of NTA and NAA codons. Among different *D. melanogaster* genes there is a correlation between degree of synonymous codon nonrandomness and levels of expression, suggesting that translational selection may be operating in *D. melanogaster*, as in prokaryotes and yeast (SHIELDS *et al.* 1988).

TABLE 1

Base utilization at position III of codons for vitellogenin and chorion genes of *D. melanogaster* and *C. capitata*

Species*	A	T	G	C	Total	% (G+C)
Drosophila						
YP1	17	70	146	207	440	80.22
YP2	19	68	151	205	443	80.36
YP3	15	65	139	202	421	80.99
s36	25	52	87	123	287	73.17
s38	34	75	64	134	307	64.49
Ceratitis						
VG1	90	113	83	151	437	53.31
VG2	82	137	81	137	421	51.78
s36	74	116	42	87	321	40.18
s38	60	106	38	78	282	41.13

* Sources of sequences: Drosophila *YP1*, *YP2* and *YP3* have EMBL nucleic acid database accession numbers V00248, J01157, and M15898, respectively. Drosophila *s36* and *s38* chorion DNA sequences are from EMBL entry X12635; Ceratitis *s36* and *s38* sequences are EMBL entries X51342 and X55886, respectively.

We compared codon usage in *D. melanogaster* and in *C. capitata* for the vitellogenin genes and for two chorion protein genes, *s36* and *s38* (KONSOLAKI *et al.* 1990; TOLIAS *et al.* 1990). A summary of the results is shown in Table 1. The *D. melanogaster* genes exhibit the deficiency of A and T in the third position which is typical for most abundantly expressed genes of this species; the bias is stronger in the vitellogenins (80% to 81% G + C in the third position) than in the chorion proteins (73% and 64% for *Ccs36* and *Ccs38*, respectively). In contrast, the vitellogenin genes of medfly show a rather even synonymous codon usage (approximately 50% G + C in the third position), while the medfly chorion genes show a slightly reversed bias (approx. 40% G + C). In all four medfly genes there is also a small bias against GTA, ATA and NCG codons, which is not observed in *D. melanogaster* (results not shown).

There are two alternative explanations for the observed difference between the two species. First, it is possible that selective pressure for specific synonymous codons is less strong in the vitellogenin and chorion genes of the medfly, because these genes are not as abundantly expressed as in *D. melanogaster*. All four genes code for proteins required at high levels during oogenesis, and *C. capitata*, with a life cycle twice as long as *D. melanogaster* and a comparable

number of egg output may have lower rates of expression of these genes than *D. melanogaster*. Alternatively, selection of synonymous codons may not be operating at all in the medfly if its effective population size is small. As has been pointed out (SHIELDS *et al.* 1988), the selection coefficients for synonymous codons are expected to be very low, and selection is possible as long as $N_e s > 1$, where N_e is the effective population size and s is the difference in selection coefficients between synonymous codons. Sequence data from medfly genes encoding weakly and highly expressed proteins may resolve this question.

Differences in codon usage patterns have been observed even among species of the genus Drosophila. In members of the Sophophora subgenus, the gene encoding alcohol dehydrogenase exhibits a more biased codon usage (similar to that of *D. melanogaster*) than in members of the Drosophila subgenus (SULLIVAN, ATKINSON and STARMER 1990). Taken together with the medfly data, these observations should serve as caution whenever phylogenetic distances are inferred from synonymous substitution rates. To minimize influences by differences in codon bias, we propose that weakly expressed genes with demonstrated low codon usage bias are used in such studies.

We thank F. C. KAFATOS for a critical reading of the manuscript. This work was supported by a U.S. Department of Agriculture grant.

LITERATURE CITED

AOTA, S., and T. IKEMURA, 1988 Diversity in G + C content in the third position of codons in vertebrate genes and its cause. Nucleic Acids Res. **14:** 6345–6355.

ASHBURNER, M., M. BODMER and F. LEMEUNIER, 1984 On the evolutionary relationships of *Drosophila melanogaster*. Dev. Genet. **4:** 295–31.

BENTON, W. D., and R. W. DAVIS, 1977 Screening of recombinant clones by hybridization to single plaques in situ. Science **196:** 180–182.

BEVERLEY, S. M., and A. C. WILSON, 1984 Molecular evolution in *Drosophila* and the higher Diptera. II. A time scale for fly evolution. J. Mol. Evol. **21:** 1–13.

BLOBEL, G., and B. DOBBERSTEIN, 1975 Transfer of proteins across membranes. J. Cell Biol. **67:** 852–862.

BODMER, M., and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature **304:** 425–430.

BOWNES, M., and B. D. HAMES, 1978 Analysis of the yolk proteins in *Drosophila melanogaster*. FEBS Lett. **96:** 327–330.

BOWNES, M., A. SHIRRAS, M. BLAIR, J. COLLINS and A. COULSON, 1988 Evidence that insect embryogenesis is regulated by ecdysteroids released from yolk proteins. Proc. Natl. Acad. Sci. USA **85:** 1554–1557.

BRENNAN, M. D., A. J. WEINER, T. J. GORALSKI and A. P. MAHOWALD, 1982 The follicle cells are a major site of vitellogenin synthesis in *Drosophila melanogaster*. Dev. Biol. **89:** 225–236.

CAVENER, D. R., 1987 Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. Nucleic Acids Res. **15:** 1353–1361.

DAYHOFF, M. O., W. C. BARKER and L. T. HUNT, 1983 Establishing homologies in protein sequences. Methods Enzymol. **91:** 524–545.

DENHARDT, D. T., 1966 A membrane-filter technique for the detection of complementary DNA. Biochem. Biophys. Res. Commun. **23:** 641–652.

FEINBERG, A. P., and B. VOGELSTEIN, 1983 A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132:** 6–13.

GARABEDIAN, M. J., M.-C. HUNG and P. C. WENSINK, 1985 Independent control elements that determine yolk protein gene expression in alternative *Drosophila* tissues. Proc. Natl. Acad. Sci. USA **82:** 1396–1400.

GARABEDIAN, M. J., B. M. SHEPHERD and P. C. WENSINK, 1986 A tissue-specific transcription enhancer from the *Drosophila* yolk protein 1 gene. Cell **45:** 859–867.

GARABEDIAN, M. J., A. D. SHIRRAS, M. BOWNES and P. C. WENSINK, 1987 The nucleotide sequence of the gene coding for *Drosophila melanogaster* yolk protein 3. Gene **55:** 1–8.

GARNIER, J., D. J. OSGUTHORPE and B. ROBSON, 1978 Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins. J. Mol. Biol. **120:** 97–120.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7074.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER and A. PAVE, 1980 Codon catalogue usage and the genome hypothesis. Nucleic Acids Res. **8:** r49–r62.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalogue usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9:** r43–r74.

HATZOPOULOS, P., and M. P. KAMBYSELLIS, 1987 Isolation and structural analysis of *Drosophila grimshawi* vitellogenin genes. Mol. Gen. Genet. **206:** 475–484.

HAYMER, D. S., J. E. ANLEITNER, M. HE, S. THANAPHUM, S. H. SAUL, J. IVY, K. HOUTCHENS and L. ARCANGELI, 1990 Actin genes in the mediterranean fruit fly *Ceratits capitata*. Genetics **125:** 155–160.

HERSKOWITZ, I., 1987 Functional inactivation of genes by dominant negative mutations. Nature **329:** 219–222.

HIGGINS, D. G., and P. M. SHARP, 1988 CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene **73:** 237–244.

HOLMES, D. S., and J. BONNER, 1973 Preparation, molecular weight, base composition, and secondary structure of giant nuclear ribonucleic acid. Biochemistry **12:** 2330–2338.

HULTMARK, D., R. KLEMENTZ and W. J. GEHRING, 1986 Translational and transcriptional control elements in the untranslated leader of the heat-shock gene *hsp22*. Cell **44:** 429–438.

HUNG, M.-C., and P. C. WENSINK, 1983 Sequence and structure conservation in yolk proteins and their genes. J. Mol. Biol. **164:** 481–492.

IKEMURA, T., 1985 Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. **2:** 13–34.

JORNVALL, H. M., M. PERSSON *and* J. JEFFREY, 1981 Alcohol and polyol dehydrogenases are both divided into two protein types, and structural properties cross-relate the different enzyme activities within each type. Proc. Natl. Acad. Sci. USA **78:** 4226–4230.

JORNVALL, H. M., H. BAHR-LINDSTROM, K. D. JANY, W. ULMER and M. FROSCHLE, 1984 Extended superfamily of short alcohol-polyol-sugar dehydrogenases: Structural similarities between glucose and ribitol dehydrogenases. FEBS Lett. **165:** 190–196.

KELLER, E., and W. A. NOON, 1985 Intron splicing: a conserved internal signal in introns of *Drosophila* pre-mRNAs. Nucleic Acids Res. **13:** 4971–4981.

KONSOLAKI, M., K. KOMITOPOULOU, P. P. TOLIAS, D. L. KING, C.

SWIMMER and F. C. KAFATOS, 1990 The chorion genes of the medfly, *Ceratits capitata*. I. Structural and regulatory conservation of the s36 gene relative to two *Drosophila* species. Nucleic Acids Res. **18:** 1731–1737.

KOZMA, R., and M. BOWNES, 1986 Yolk protein induction in males of several *Drosophila* species. Insect. Biochem. **16:** 263–271.

MALACRIDA, A., G. GASPERI, G. F. BISCALDI and R. MILANI, 1986 Persistence of linkage relationships among enzyme loci in some Dipteran species. Atti Assoc. Genet. Ital. **31:** 121–122.

MANIATIS, T., E. F. FRITCH and J. SAMBROOK, 1982 *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

MINTZAS, A. C., G. CHRYSANTHIS, C. CHRISTODOULOU and V. J. MARMARAS, 1983 Translation of the mRNAs coding for the major haemolymph proteins of *Ceratits capitata* in a cell-free system: comparison of the translatable mRNA levels to the respective biosynthetic levels of the proteins in the fat body during development. Dev. Biol. **95:** 492–496.

MOUNT, S. M., 1982 A catalogue of splice junction sequences. Nucleic Acids Res. **10:** 459–472.

MULLER, H. J., 1940 Bearings of the *Drosophila* work on systematics, pp 185–268 in *The New Systematics*, edited by J. HUXLEY, Clarendon Press, Oxford.

NARDELLI, D., S. GERBER-HUBER, F. D. VAN HET SCHIP, M. GRUBER, G. AB and W. WAHLI, 1987 Vertebrate and nematode genes coding for yolk proteins are derived from a common ancestor. Biochemistry **26:** 6397–6402.

PERSSON, B., G. BENGTSSON-OLIVECRONA, S. ENERBACK, T. OLIVECRONA and H. JORNVALL, 1989 Structural features of lipoprotein lipase. Lipase family relationships, binding interactions, non-equivalence of lipase cofactors, vitellogenin similarities and functional subdivision of lipoprotein lipase. Eur. J. Biochem. **179:** 39–45.

POSTLETHWAIT, J. H., M. BOWNES and T. JOWETT, 1980 Sexual phenotype and vitellogenin synthesis in *Drosophila melanogaster*. Dev. Biol. **79:** 379–387.

PROUDFOOT, N. J. and G. G. BROWNLEE, 1976 3′ Non-coding region sequences in eukaryotic messenger RNA. Nature **263:** 211–214.

RIDDIHOUGH, G., and H. R. B. PELHAM, 1987 An ecdysone response element in the *Drosophila hsp70* promoter. EMBO J. **6:** 3729–3734.

RINA, M. D., and A. C. MINTZAS, 1987 Two vitellins-vitellogenins of the Mediterranean fruit fly *Ceratitis capitata*: a comparative biochemical and immunological study. Comp. Biochem. Physiol. **86B:** 801–808.

RINA, M. D., and A. C. MINTZAS, 1988 Biosynthesis and regulation of two vitellogenins in the fat body and ovaries of *Ceratitis capitata* (Diptera). Roux's Arch. Dev. Biol. **197:** 167–174.

SHARP, P. M., and W-H. LI, 1986 On the rate of DNA sequence evolution in *Drosophila*. Nucleic Acids Res. **14:** 7737–7749.

SHARP, P. M., E. COWE, D. G. HIGGINS, D. C. SHIELDS, K. H. WOLFE and F. WRIGHT, 1988 Codon usage patterns in *Escherichia coli, Bacillus subtilis, Sacchromyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. **16:** 8207–8211.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS and F. WRIGHT, 1988 "Silent" sites in *Drosophila* genes are not neutral: Evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SPIETH, J., K. DENISON, S. KIRTLAND, J. CANE and T. BLUMENTHAL, 1985 The *C. elegans* vitellogenin genes: short sequence repeats in the promoter regions and homology to the vertebrate genes. Nucl. Acids Res. **13:** 5283–5295.

STADEN, R., 1984 Graphic methods to determine the function of nucleic acid sequences. Nucleic Acids Res. **12:** 521–538.

SULLIVAN, D. T., P. W. ATKINSON and W. T. STARMER, 1990 Molecular evolution of alcohol dehydrogenase genes in the genus *Drosophila*. Evol. Biol. **24:**107–147.

TEEM, J. L., N. A. ABOVICH, N. F. KAUFER, W. F. SCHWINDINGER, J. R. WARNER, A. LEVY, J. WOOLFORD, R. J. LEER, M. M. C. VAN RAAMSDONK-DUIN, W. H. MAGER, R. J. PLANTA, L. SCHULTZ, J. D. FRIESEN, H. FRIED and M. ROSBASH, 1984 A comparison of yeast ribosomal protein gene DNA sequences. Nucleic Acids Res. **12:** 8295–8312.

TOLIAS, P. P., M. KONSOLAKI, K. KOMITOPOULOU and F. C. KAFATOS, 1990 The chorion genes of the medfly, *Caratits capitata*. II. Characterization of three novel cDNA clones obtained by differential screening of an ovarian library. Dev. Biol. **140:** 105–112.

WALLACE, R. B., M. J. JOHNSON, S. Y. SUGGS, K. MIYOSHI, R. BHATT and K. ITAKURA, 1981 A set of synthetic oligodeoxyribonucleotide primers for DNA sequencing in the plasmid vector pBR322. Gene **16:** 21–26.

WARREN, T. J., and A. P. MAHOWALD, 1979 Isolation and partial chemical characterization of the three major egg yolk polypeptides from *Drosophila melanogaster*. Dev. Biol. **68:** 130–139.

YAN, Y. L., C. J. KUNERT and J. H. POSTLETHWAIT, 1987 Sequence homologies among the three yolk polypeptide (*Yp*) genes in *Drosophila melanogaster*. Nucleic Acids Res. **15:** 67–85.

ZACHAROPOULOU, A., 1990 Polytene chromosome maps in the medfly *Ceratitis capitata*. Genome **33:** 184–197.

Communicating editor: W. M. GELBART

Pergamon

0965-1748(95)00037-2

Insect Biochem. Molec. Biol. Vol. 25, No. 8, pp. 939–958, 1995

0965-1748/95 $9.50 + 0.00

Analysis of a Vitellogenin Gene of the Mosquito, *Aedes aegypti* and Comparisons to Vitellogenins from Other Organisms

PATRICIA ROMANS,* ZHIJIAN TU,† ZHAOXI KE,†‡ HENRY H. HAGEDORN†§

Received 22 April 1993; revised and accepted 12 April 1995

A genomic clone of the *Aedes aegypti* vitellogenin A1 gene was sequenced¶ including 2015 bp of 5′ untranscribed sequence, 6369 bp of open reading frame interrupted by two introns, and a short 3′ untranslated region. Primer extension was used to identify the transcription initiation site. The amino termini of the large and small subunits were located by N-terminal sequencing of vitellin purified from eggs. The length of the signal sequence and the position of the cleavage site between the two subunits were also determined. Three sequential imperfect repeats were found near the beginning of the small subunit. The sequence of the coding region appears to be polymorphic. Comparison of the signal sequences of seven insect vitellogenin genes revealed several conserved leucines, and a conserved position of an intron. However, the signal sequences are not conserved between these genes and the yolk protein genes of Cyclorraphid Dipteran insects. The cleavage sites between the small and large subunits in the vitellogenins of the mosquito, *A. aegypti*, sawfly, *Athalia rosae*, boll weevil, *Anthonomus grandis*, and silkworm, *Bombyx mori* are flanked by sequences rich in serine. Pairwise dot matrix analysis at the protein level showed that the mosquito, boll weevil and silkworm vitellogenins are significantly related with approx. 50% similarity. One region of the three insect vitellogenin genes, near the N-terminal of the large subunit, showed the highest levels of similarity, from 57.5 to 64.4%. The position of cysteines in insect vitellogenins is conserved, particularly in the C-terminus of the large subunit. Dot matrix comparison of the mosquito vitellogenin with that of *Xenopus laevis* and *Caenorhabditis elegans* showed much lower, but still significant degrees of relationship. Pairwise comparisons of the mosquito vitellogenin and the *Drosophila melanogaster* yolk proteins did not show significant similarities. Potential regulatory regions in the mosquito VgA1 gene were identified by comparison to regulatory elements known from other organisms, especially *D. melanogaster*, which could provide useful information for further functional analysis.

Aedes aegypti Hormone response element Sequence comparison Vitellogenin gene Ecdysone

INTRODUCTION

The major egg proteins have been the subject of intense study for many years. Before the technological revolutions that allowed the analysis of macromolecules from small samples, the massive amounts of vitellogenin and ovalbumin produced during egg development, and their concentration in the egg, made these proteins the molecules of choice for many kinds of biochemical investigations. A considerable amount is known about the biochemistry and physiology of these proteins which forms the basis for the current intensive investigations of their molecular biology.

Evidence for the distant evolutionary relatedness of vitellogenin genes of vertebrates and nematodes exists in their primary structure, and in the positions of introns (Nardelli *et al.*, 1987; Spieth *et al.*, 1991; Trewitt *et al.*, 1992). However, comparatively little is known about the relationships among insect vitellogenin genes, and the relationships between insect vitellogenin genes and those of other organisms. Genomic sequences are now available for the vitellogenin genes of the boll weevil, *Anthonomus grandis* (Trewitt *et al.*, 1992), the silkworm, *Bombyx mori* (Yano *et al.*, 1994b), the mosquito, *Anopheles gambiae* (P. Romans, unpublished), and the yolk proteins of the fruit flies, *Drosophila melanogaster* (Hung and Wensink, 1983; Garabedian *et al.*, 1987)

*Department of Zoology, University of Toronto, Toronto, Ontario, Canada M5S 1A1.
†Department of Entomology and Center for Insect Science, University of Arizona, Tucson, AZ 85721, U.S.A.
‡Centers for Disease Control, Atlanta, GA, U.S.A.
§Author for correspondence.
¶Genbank accession number L41842.

Exhibit C

and *Ceratitis capitata* (Rina and Savakis, 1991), and the blowfly, *Calliphora erythrocephala* (Martinez and Bownes, 1994). cDNA sequences have been obtained for the mosquito, *Aedes aegypti*, vitellogenin gene (Chen *et al.*, 1994) and the silkworm (Yano *et al.*, 1994a). Partial sequences of vitellogenin genes for the sawfly, *Athalia rosae* (Kageyama *et al.*, 1994); locust, *Locusta migratoria* (Locke *et al.*, 1987); and Gypsy moth, *Lymantria dispar* (Hiremath *et al.*, 1994), are also available.

Among the Diptera there are major differences in the molecules used as vitellogenins. In the lower Diptera, which include the mosquitoes, vitellogenins resemble the molecules seen in other insects and vertebrates where they are composed of large glycolipoproteins with large (>120 k) and small (~55 k) subunits (Kunkel and Nordin, 1985). In the Cyclorraphid Diptera, such as the fruit fly, *D. melanogaster*, the vitellogenins (appropriately called yolk proteins) are quite different from other animal vitellogenins, being composed of several small (~45 k) molecules related to mammalian tricylglycerol lipase (Baker, 1988; Bownes *et al.*, 1988; Terpstra and AB, 1988). A comparative study of the vitellogenin and yolk protein genes could provide insight into the evolution of these molecules.

A major focus of current work on vertebrate and invertebrate vitellogenins concerns the hormonal control of gene expression. This work provides some of the most detailed models for endocrine regulation of gene expression. Estrogen regulates vitellogenin gene expression in the vertebrates (Corthesy *et al.*, 1990). Among the insects, juvenile hormone appears to be the most important regulatory hormone (Koeppe *et al.*, 1985), with the exception of the Diptera where 20-hydroxyecdysone substitutes for juvenile hormone in some species (Hagedorn, 1985, 1994).

In the mosquito, *A. aegypti*, vitellogenin is synthesized in the fat body after a blood meal under the control of 20-hydroxyecdysone, however, juvenile hormone also has effects on vitellogenin synthesis prior to the blood meal in ways that are not well understood (Borovsky *et al.*, 1985; Racioppi *et al.*, 1986; Martinez and Hagedorn, 1987; Hagedorn, 1994). One of our goals is to use molecular techniques to study the hormonal control of vitellogenin gene expression. Thus, a comparison of the regulatory regions of mosquito genes with other genes controlled by steroid hormones could identify potential hormone response elements for future functional analysis.

This paper presents the genomic sequence analysis of an expressed vitellogenin gene of the mosquito, *A. aegypti*.

MATERIALS AND METHODS

Animals

A. aegypti derived from the NIH-Rockefeller strain were reared as described by Shapiro and Hagedorn (1982). Three- to four-day-old females were fed on warmed pig blood (37°C) through a Parafilm membrane.

Purification of vitellin from eggs

Ovaries containing vitellogenic oocytes were dissected from females fed 24 h earlier and immediately frozen. Ovaries were homogenized at 0°C in a 0.5 M Tris–PO_4 buffer, pH 8.0 containing 0.4 M NaCl and 0.1 mM DFP as a protease inhibitor, and the homogenate was centrifuged for 5 min at maximum speed in a Beckman (Palo Alto, Calif.) microcentrifuge at 4°C. The supernatant was dialyzed against a 0.05 M Tris–HCl buffer, pH 8.0 containing 0.05 M NaCl and bound to a DEAE (Whatman-52) column. Vitellin was eluted with a 0.05–0.5 M gradient of NaCl at 0.25 M.

N-terminal amino acid sequence

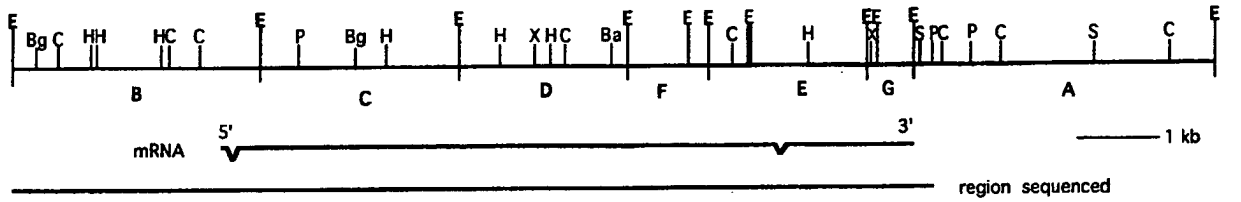
Vitellin protein prepared as described above was separated by PAGE and blotted onto PVDF membrane. Regions containing the large and small subunits were cut from the membrane and the N-terminal amino acid sequence was determined by automated Edman degradation at the University of Arizona Biotechnology Core Facility using an Applied Biosystems 477A Protein sequencer interfaced with a 120A HPLC (C-18 PTH, reverse-phase chromatography) analyzer to determine phenylthiohydantoin (PTH) amino acids.

Nucleotide sequencing

DNA sequencing was performed on EcoR I, Hind III, EcoR I-Hind III and Hind III-Sal I subclones of the VgA1 lambda clone in pBluescript II KS$^-$ (Stratagene). Sequencing primers were T3, T7 (Promega) and several synthetic oligonucleotides based on previously obtained sequence prepared at the University of Arizona Division of Biotechnology, or University of Toronto Zoology Molecular Core Facility. Double stranded sequencing template DNA was purified over Qiagen columns. Sequence was obtained manually using α ^{35}S-dATP, >1000 Ci/m mol (Amersham), a Sequenase version 2.0 kit (United States Biochemical Corp., and Amersham) and standard 6% acrylamide urea gels, or by an automatic DNA sequencer (model 373A, Applied Biosystems Int., Forster City, Calif.) at either the Core Facility for Protein/DNA Chemistry, Queen's University, or the Division of Biotechnology, University of Arizona.

Polymerase chain reaction (PCR)

PCR reactions were carried out to double-check the sequences of two regions in the B and C fragments (Fig. 1), using either the genomic DNA (0.45 μg template per reaction) of *A. aegypti* or the cloned plasmids (0.06–0.6 μg template per reaction) as template. PCR reactions were carried out in a total volume of 100 μl with 10 μl of 10 × buffer (200 mM Tris–HCl, pH 8.4, 500 mM KCl) 1.6 mM of Mg^{2+}, 0.1 mM dNTPs, 0.5 μM of each primer, and 2.5–5.0 units of Taq polymerase. Templates were denatured at 94°C for 1 min.



Ba=BamHI, Bg=BglII, C=ClaI, E=EcoRI, H=HindIII, P=PstI, S=SalI, X=XhoI

FIGURE 1. Restriction enzyme map of the genomic clone in phage lambda containing the VgA1 gene of *A. aegypti*. The clone is divided into fragments (A to G) defined by EcoR1 sites as described by Gemmill *et al.* (1986). Additional EcoR I fragments were identified during remapping and sequencing. The location of the transcribed region (mRNA) is indicated below the restriction map. Also shown is the region sequenced.

Extension was carried out at 72°C and the extension time was calculated as 1 kb/min. Annealing temperature was 54°C. PCR products were sequenced by the Division of Biotechnology of the University of Arizona using synthetic primers and the automatic sequencer.

RNA preparation

Total RNA was isolated by modification of the single-step RNA isolation method (Chomczynski and Sacchi, 1987). Adult female *A. aegypti* were frozen at −80°C 24–30 h after blood meal, quickly shaken in a cold Erlenmeyer flask and sieved to remove the heads which contained a pigment that interfered with some of the subsequent steps. The thoraces and abdomens were then ground in a mortar and pestle on dry ice. Five ml of denaturing solution containing 4 M guanidinium thiocyanate was added and after mixing thoroughly, 0.5 ml samples were placed into 1.5 ml polypropylene tubes. Water-saturated phenol–chloroform extraction and ethanol precipitation were carried out essentially as described by Chomczynski and Sacchi (1987). RNA was purified by (1) re-extracting with phenol-chloroform (pH 7.2–7.4) after dissolving the RNA pellet in DEPC-treated 0.5% SDS, or (2) using an RNaid plus kit (BIO 101, Vista, Calif.) following the manufacturer's instructions.

Primer extension analysis

A 21-mer synthetic oligonucleotide complementary to a region immediately upstream of the translation initiation site (Fig. 2) was synthesized as above, and purified by polyacrylamide gel electrophoresis. 100 ng of primer was end-labeled using 10 units of T4 polynucleotide kinase in a 10 μl reaction containing 3 μl (γ-^{32}P) ATP (3000 Ci/mmol, 10 mCi/ml) in 50 mM Tris–HCl buffer (pH 7.5) plus 1 mM $MgCl_2$, 5 mM DTT and 0.1 mM spermidine. After a 10 min incubation at 37°C the sample was heated to 90°C for 2 min to inactivate the kinase. The final concentration was adjusted to 100 fmol/μl and stored at −20°C.

In the reaction, 100 fmol of ^{32}P-labeled primer and 10 μg of total RNA were mixed in 100 mM Tris–HCl (pH 8.3), 50 mM KCl, 10 mM $MgCl_2$, 10 mM DTT, 1 mM of each dNTP and 1 mM spermidine in a volume of 10 μl. The primer was annealed to the RNA by heating the tube to 58°C for 20 min, and allowing the sample to cool at room temperature for 10 min. One unit of AMV reverse transcriptase was added in 2.8 mM $Na_4P_2O_7$ (total volume, 20 μl). The reaction was incubated at 42°C for 30 min. An equal volume of loading buffer (98% formamide, 10 mM EDTA, 0.1% Xylene cyanol, 0.1% bromophenol blue) was added and heated at 90°C for 10 min immediately before loading onto an 8% polyacrylamide gel containing 7 M urea. The gel was dried and submitted to autoradiography.

Ribonuclease protection assay

DNA of the subloned B fragment of VgA1 (Fig. 1) was digested with Cla I and treated with proteinase K (100 mg/ml) for 1 h at 37°C followed by a phenol–chloroform extraction and precipitation with ethanol. The DNA was dissolved in DEPC-treated water at a concentration of 1 μg/μl. A 634 base antisense RNA probe, labeled with (α-^{32}P)-UTP, was transcribed from the T3 promotor using an RNA transcription kit (Stratagene, La Jolla, Calif.). The DNA template was then digested by adding RNase-free DNase I (Boehringer Mannheim, Indianapolis, Ind.) at 37°C for 15 min. Unincorporated nucleotides were removed by passing the reaction over a Quick-spin G-50 column (Boehringer Mannheim). The eluted RNA was precipitated with ethanol, and dissolved in 100 μl of hybridization buffer (40 mM PIPES, pH 6.4, 1 mM EDTA, 0.4 M NaCl, 80% formamide). This RNA probe was mixed with 10 μg total RNA from blood-fed female *A. aegypti*, denatured at 85°C for 10 min, and transferred to a water bath at 50°C overnight. The hybridization mixture was incubated with RNase digestion mixture (300 mM NaCl, 10 mM Tris–Cl, pH 7.4, 5 mM EDTA, 40 μg/ml RNase A1, 2 μg/ml RNase T1) at 30°C for 1 h. 20 μl of 10% SDS and 10 μl of a 10 mg/ml solution of proteinase K were added and incubated at 37°C for 30 min. The RNA was extracted with phenol–chloroform, precipitated with ethanol, resuspended in formamide loading buffer, heated at 95°C for 5 min and then analyzed on a 6% polyacrylamide gel containing 7 M urea. The gel was dried and examined by autoradiography.

```
gaattccaccaccaggcagtgctagtgtgcatgaactgaaagatggcgtcttcggtaaag -1956
ttgttaacatgttcactagtgccacctggtggcaaaatcttgaatttcaacaatgatagc -1896
ctttcacttattgaacaagtttgtcgaagacgccatctttataagtagtcaagatttgga -1836
gattacggcaaaccaaaagattaatattcactagcgccatctgatggctaaatttcgaat -1776
ttatcggctagattaatgatagatctaaagctgctgaaaaactttgccgaatacgtcatc -1716
tttctgagagaccaggatactgagttatttttaaaacaaaggttccgtgctcactaatgc -1656
cgcatggtggctaaatggcttttatactgaaaacgcttgagctttgcctaacacgtcgtt -1596
tttttaggtagtgaggaacctgatataacagaataaaatcgatacaccagcgccgcctgg -1536
attccgcatatctattaccaccgcatgttagccctttatccaaataaaaaaaattactga -1476
acatcatgactctctatgttgatgacttttcaataagcatttataataacggtcgtcaat -1416
ctcaatttcattcgtatatataaccatcgtagtacgaatggtcataaaaagaggttgaag -1356
tgtgtttggattctccccagacaaaaaatcgaaatgaaaaactgaaacccattttgacaa -1296
tcgtcggaaaggtctttctgtttagttcactgtaacaaaatgcaatccaaagatatgaaa -1236
gctttgaaaacggcgaaaagttattatatctaccgttttattacggaaagcttcttattc -1176
cggacactctactttgtatgagaaacatttcatatgagatgtttcaatttttgctgttca -1116
aaagttcacacttttgaggttcattttaatcaaaattttctatagatatctatgaaaatt -1056
tataatactcaactgctttagacgcctctttagaggcttgccaatttcaaataatgattt -996
gagcttgtctttttttatgattttcatgagctgtccggaatttgaatcaaagtgtccggaa -936
tatggggcaaaagtaaggaagcgtccggaataagaatcgtgaaaagtccacacatttcta -876
tttattgaaaatcattcatgtatcagaaacaaaatttttcacccaccattcgaaagttaag -816
tgttttgaaggcttagagccgaggaatcatacagattgatttatttttatatgcttcctga -756
tgggttatacttcattgaggccttaagtgtccgtaatatgaatcaaaacggtacatatgt -696
catggcaacattttccgaaacctggagataaattacagagcccactcagtatcgcagatg -636
ccaacagtgtggctatggtaccaaacatcgtgcaaaagtttgtcattttactcaataact -576
gacttgtataattagtttaagcttattcaaatgctttttacaagctgatgatttcaattt -516
gcctgtgaaaattctgaactgttgcggcaaatgaaatgcaatcgatagaaacaaataatg -456
tagcaatagcaaaaaataatcatttttttgcttatcttactatcttcaattcacatctgta -396
gtctcaattgaataatctggaatccattgcaagctaagtaaattcacgtgtgacctagcg -336
ggaggccaatggtcgagtgaatctttatttcttgaatggcagaaacgatgccatgaatca -276
```

Fig. 2—*legend on p. 948.*

```
    aatccaggatgcgaaaccgatgcacaagaataccaatacgaaaactaatcccaatacaac -216

    gatcaccaggtgcatcgatgcgaccgcgattactttgttaccatctgttgtgctgcggtt -156

    atcgcttttcgattagaaggcgaacgctgaaccgatcgtatgcttatcatcgcgaacga -96

    aaggatgctgtgaatcactgctgatgggggcaaaaatctacgaaaatgtaagcaatcact -36

    ttcaaatataaaaccttccaatggccacagacggtatcacttctcgttttggtttcaacg 25
         TATA box                          +1
    agaggaggagaacacacaatcggaacagctgccggatacttgaagacaagATGCTAGCGA 85
 1                                                    M  L  A  K

    AACTACTTCTTCTCGCTTTGGgtaagtgctcccgggaatgttcgcctcaaacttcgaata 145
 5    L  L  L  L  A  L  A

    ctattctgttctttcctttcgacttccacagCGGGGCTCACTGCTGCCTACCAATACGAG 205
12                                   G  L  T  A  A  Y  Q  Y  E
                                                    ==========
    AACTCGTTCAAGGGCTACAATCCTGGCTATAAGGGCTACGATGCTGGCTACAAGGGTTAC 265
21  N  S  F  K  G  Y  N  P  G  Y  K  G  Y  D  A  G  Y  K  G  Y
    ================================
                                 ****************************  *****
    GGCTACGATGCTGGCTACAAAGGCTACGGATACGATGCTGGTTACAAATACAACAACCAA 325
41  G  Y  D  A  G  Y  K  G  Y  G  Y  D  A  G  Y  K  Y  N  N  Q
    ******************  **************************
    GGCTACAGTTACAAGAACGGTTTCGAATATGGATATCAGAACGCCTACCAGGCTGCTTTC 385
61  G  Y  S  Y  K  N  G  F  E  Y  G  Y  Q  N  A  Y  Q  A  A  F

    TATAAGCACCGTCCAAACGTAACCGAATTCGAGTTCAGCTCATGGATGCCGAACTACGAG 445
81  Y  K  H  R  P  N  V  T  E  F  E  F  S  S  W  M  P  N  Y  E

    TACGTCTACAATGTGACGTCCAAGACCATGACCGCTCTGGCGGAATTGGACGATCAGTGG 505
101 Y  V  Y  N  V  T  S  K  T  M  T  A  L  A  E  L  D  D  Q  W

    ACTGGTGTTTTCACCCGTGCCTACCTGGTCATCCGTCCCAAGAGCCGTGACTACGTCGTG 565
121 T  G  V  F  T  R  A  Y  L  V  I  R  P  K  S  R  D  Y  V  V

    GCTTACGTCAAGCAGCCAGAATACGCTGTCTTCAACGAACGCCTGCCACACGGATATGCT 625
141 A  Y  V  K  Q  P  E  Y  A  V  F  N  E  R  L  P  H  G  Y  A

    ACCAAGTTCTATCACGATATGTTCAAGTTCCAACCAATGCCAATGAGCAGCAAGCCATTC 685
161 T  K  F  Y  H  D  M  F  K  F  Q  P  M  P  M  S  S  K  P  F

    GGAATCCGTTACCATAAGGGCGCCATCAAGGGTCTGTACGTCGAGAAAACCATTCCCAAC 745
181 G  I  R  Y  H  K  G  A  I  K  G  L  Y  V  E  K  T  I  P  N

    AATGAAGTCAACATCCTGAAGGCTTGGATCAGTCAGCTGCAGGTTGATACCCGTGGAGCC 805
201 N  E  V  N  I  L  K  A  W  I  S  Q  L  Q  V  D  T  R  G  A

    AACTTGATGCATTCCAGCAAGCCCATCCATCCTTCCAAGAATGAGTGGAACGGCCATTAC 865
221 N  L  M  H  S  S  K  P  I  H  P  S  K  N  E  W  N  G  H  Y

    AAGGTTATGGAGCCATTGGTTACCGGAGAATGTGAAACCCATTACGATGTCAACCTGATC 925
241 K  V  M  E  P  L  V  T  G  E  C  E  T  H  Y  D  V  N  L  I

    CCAGCCTACATGATCCAAGCTCACAAACAGTGGGTTCCTCAAGGACAACTCCGTGGTGAA 985
261 P  A  Y  M  I  Q  A  H  K  Q  W  V  P  Q  G  Q  L  R  G  E
```

Fig. 2 (continued)—*legend on p. 948.*

```
    GATGATCAGTTCATTCAAGTCACCAAGACTCAGAACTTTGACCGTTGCGATCAACGCATG 1045
281 D  D  Q  F  I  Q  V  T  K  T  Q  N  F  D  R  C  D  Q  R  M

    GGTTACCACTTTGGATTCACCGGATACAGCGATTTCAGACCCAACACCAACCAAATGGGA 1105
301 G  Y  H  F  G  F  T  G  Y  S  D  F  R  P  N  T  N  Q  M  G

    AATGTTGCCTCCAAGTCTTTGGTTTCATACATGTATTTGACTGGAAACTGGTACAACTTC 1165
321 N  V  A  S  K  S  L  V  S  Y  M  Y  L  T  G  N  W  Y  N  F

    ACCATCCAATCGTCCAGCATGATCAACAAGGTGGCTATCGCTCCTTCCCTAGTGAACAAG 1225
341 T  I  Q  S  S  S  M  I  N  K  V  A  I  A  P  S  L  V  N  K

    GAACCAGCTCTCGTGTACGCTCAGGTTAACATGACCCTCAACGATGTCCACCCTTACGAT 1285
361 E  P  A  L  V  Y  A  Q  V  N  M  T  L  N  D  V  H  P  Y  D

    AAGGTCCCAATGGGCCCAGCCGAAGATCTGAAGGTGTTCGTCGATTTGGTTTACAGCTAC 1345
381 K  V  P  M  G  P  A  E  D  L  K  V  F  V  D  L  V  Y  S  Y

    AACATGCCAAGTGATAAGAAGAACTACGTTCGTCCTGGCAACGAAACCTCCTCCTCCTCA 1405
401 N  M  P  S  D  K  K  N  Y  V  R  P  G  N  E  T  S  S  S  S

    TCGTCATCTTCTTCATCGTCCTCCTCCTCATCGGAATCTAGCTCGTCCAGCTCTGAATCT 1465
421 S  S  S  S  S  S  S  S  S  S  E  S  S  S  S  S  S  E  S

    GTGGAAAACCCCAAGATCAGCCCCGTTGAGCAGTACAAGCCTCTGCTGGACAAAGTTGAG 1525
441 V  E  N  P  K  I  S  P  V  E  Q  Y  K  P  L  L  D  K  V  E

    AAGCGTGGAAACCGCTACCGTCGTGATCTGAATGCCATCAAGGAAAAGAAGTACTACGAA 1585
461 K  R  G  N  R  Y  R  R  D  L  N  A  I  K  E  K  K  Y  Y  E
                              ==================================
    GCTTACAAAATGGATCAGTACCGTCTGCACCGTTTGAACGATACTTCATCCGACTCCAGC 1645
481 A  Y  K  M  D  Q  Y  R  L  H  R  L  N  D  T  S  S  D  S  S
    =======
    AGCTCTGATTCCTCATCATCCAGCTCGTCGGAATCCAAGGAACACCGCAACGGCACTTCT 1705
501 S  S  D  S  S  S  S  S  S  S  E  S  K  E  H  R  N  G  T  S

    TCCTATTCCAGCTCCTCATCGTCCTCTTCTTCATCGTCCTCTTCTGAGTCATCCTCCTAC 1765
521 S  Y  S  S  S  S  S  S  S  S  S  S  S  S  S  E  S  S  S  Y

    TCATCCTCTTCTTCCTCTTCCTCGGAGTCCTACTCCATTAGCAGCGAAGAGTACTACTAC 1825
541 S  S  S  S  S  S  S  S  E  S  Y  S  I  S  S  E  E  Y  Y  Y

    CAACCAACACCAGCTAACTTCAGCTATGCTCCCGAAGCTCCGTTCCTGCCATTCTTCACC 1885
561 Q  P  T  P  A  N  F  S  Y  A  P  E  A  P  F  L  P  F  F  T

    GGATATAAGGGATACAACATCTTCTACGCACGCAATGTTGATGCCATTCGCTCAGTCGGC 1945
581 G  Y  K  G  Y  N  I  F  Y  A  R  N  V  D  A  I  R  S  V  G

    AAACTTGTTGAGGAAATCGCCAGCGATCTGGAAAACCCATCCAACTTGCCCAAAGCCAAC 2005
601 K  L  V  E  E  I  A  S  D  L  E  N  P  S  N  L  P  K  A  N

    ACCATGAGCAAGTTCAACATTCTGACCCGTGCTATCAGAGCCATGGGATACGAAGACATT 2065
621 T  M  S  K  F  N  I  L  T  R  A  I  R  A  M  G  Y  E  D  I

    TACGAGCTGGCCCAGAAATACTTTGTTTCGCAGAAAGAACGTCAAGTCGCTCAGTTTTCC 2125
641 Y  E  L  A  Q  K  Y  F  V  S  Q  K  E  R  Q  V  A  Q  F  S

    GACAAAAAATTCAGCAAGCGCGTTGACGCTTGGGTTACCCTCCGTGATGCTGTTGCTGAA 2185
```

Fig. 2 (continued)—*legend on p. 948.*

```
661 D  K  K  F  S  K  R  V  D  A  W  V  T  L  R  D  A  V  A  E

    GCCGGAACCCCATCCGCTTTCAAATTGATTTTCGATTTCATCAAGGAAAAGAAACTGCGT 2245
681 A  G  T  P  S  A  F  K  L  I  F  D  F  I  K  E  K  K  L  R

    GGATACGAGGCTGCCACCGTGATTGCTTCCTTGGCCCAATCTATCCGCTACCCAACCGAG 2305
701 G  Y  E  A  A  T  V  I  A  S  L  A  Q  S  I  R  Y  P  T  E

    CATCTGCTGCACGAATTCTTCCTCCTGGTTACCAGCGATGTTGTATTGCACCAGGAATAC 2365
721 H  L  L  H  E  F  F  L  L  V  T  S  D  V  V  L  H  Q  E  Y

    TTGAATGCCACCGCTCTGTTCGCTTACTCCAACTTCGTCAACCAAGCTCATGTTAGCAAC 2425
741 L  N  A  T  A  L  F  A  Y  S  N  F  V  N  Q  A  H  V  S  N

    CGTTCAGCTTACAACTACTATCCAGTATTCAGTTTTGGTCGCCTGGCTGATGCCGACTAC 2485
761 R  S  A  Y  N  Y  Y  P  V  F  S  F  G  R  L  A  D  A  D  Y

    AAGATCATCGAACACAAGATCGTCCCATGGTTCGCTCACCAGCTCCGTGAAGCCGTTAAC 2545
781 K  I  I  E  H  K  I  V  P  W  F  A  H  Q  L  R  E  A  V  N

    GAAGGAGACAGTGTAAAGATCCAGGTCTACATCCGTTCCCTCGGAAACCTTGGACATCCA 2605
801 E  G  D  S  V  K  I  Q  V  Y  I  R  S  L  G  N  L  G  H  P

    CAAATCCTGTCGGTATTCGAGCCATACCTGGAGGGTACCATTCAGATCACTGACTTCCAG 2665
821 Q  I  L  S  V  F  E  P  Y  L  E  G  T  I  Q  I  T  D  F  Q

    CGCTTGGCCATTATGGTCGCTTTGGACAATCTGGTTATCTACTACCCAAGCTTGGCCCGT 2725
841 R  L  A  I  M  V  A  L  D  N  L  V  I  Y  Y  P  S  L  A  R

    TCGGTGCTTTACCGTGCCTACCAAAACACTGCCGATGTCCATGAAGTTCGTTGTGCTGCC 2785
861 S  V  L  Y  R  A  Y  Q  N  T  A  D  V  H  E  V  R  C  A  A

    GTTCATTTGTTGATGCGCACCGACCCACCAGCTGATATGCTGCAACGTATGGCCGAGTTC 2845
881 V  H  L  L  M  R  T  D  P  P  A  D  M  L  Q  R  M  A  E  F

    ACTCACCACGACCCAAGGCTCTACGTCCGCGCTGCCGTCAAATCCGCCATTGAAACTGCT 2905
901 T  H  H  D  P  R  L  Y  V  R  A  A  V  K  S  A  I  E  T  A

    GCCTTGGCTGACGACTACGACGAAGACAGCAAGTTGGCCCTTAATGCTAAGGCTGCCATT 2965
921 A  L  A  D  D  Y  D  E  D  S  K  L  A  L  N  A  K  A  A  I

    AACTTCCTGAACCCAGAAGACGTCAGCATTCAGTACTCCTTCAATCACATCCGTGACTAC 3025
941 N  F  L  N  P  E  D  V  S  I  Q  Y  S  F  N  H  I  R  D  Y

    GCTTTGGAAAACCTCGAGCTTTCCTACCGTCTGCACTACGGAGAAATCGCTTCCAATGAC 3085
961 A  L  E  N  L  E  L  S  Y  R  L  H  Y  G  E  I  A  S  N  D

    CATCGCTACCCAAGTGGACTGTTCTATCATCTGCGCCAAAACTTCGGAGGATTCAAGAAA 3145
981 H  R  Y  P  S  G  L  F  Y  H  L  R  Q  N  F  G  G  F  K  K

    TACACCTCGTTCTACTATCTGGTTTCGAGCATGGAAGCTTTCTTCGATATTTTCAAGAAG 3205
1001 Y  T  S  F  Y  Y  L  V  S  S  M  E  A  F  F  D  I  F  K  K

    CAATACAACACCAAGTACTTCGCTGATTATTACAAATCTGCCGACTACAGCACTAACTAC 3265
1021 Q  Y  N  T  K  Y  F  A  D  Y  Y  K  S  A  D  Y  S  T  N  Y

    TACAACTTTGACAAATACTCCAAGTACTACAAGCAGTACTACTACAGCAAGGACAGCGAA 3325
1041 Y  N  F  D  K  Y  S  K  Y  Y  K  Q  Y  Y  Y  S  K  D  S  E
```

Fig. 2 (continued)—*legend on p. 948.*

```
     TACTACCAGAAGTTCTACGGACAGAAGAAGGATTACTATAACGATAAGGAGCCATTCAAG 3385
1061 Y  Y  Q  K  F  Y  G  Q  K  K  D  Y  Y  N  D  K  E  P  F  K

     TTCACTGCACCACGCATTGCCAAGCTGCTGAACATCGATGCTGAAGAGGCTGAGCAACTG 3445
1081 F  T  A  P  R  I  A  K  L  L  N  I  D  A  E  E  A  E  Q  L

     GAAGGACAACTGTTGTTCAAACTGTTCAACGGATACTTCTTCACCGCTTTCGATAACCAA 3505
1101 E  G  Q  L  L  F  K  L  F  N  G  Y  F  F  T  A  F  D  N  Q

     ACCATCGAAAACCTCCCACATAAGATGAGACATCTTTTCGAAAATTTGGAAGATGGCTAC 3565
1121 T  I  E  N  L  P  H  K  M  R  H  L  F  E  N  L  E  D  G  Y

     GCTTTTGACGTTACCAAATTCTACCAACAACAGGATGTTGTTCTGGCCTGGCCTTTGGCT 3625
1141 A  F  D  V  T  K  F  Y  Q  Q  Q  D  V  V  L  A  W  P  L  A

     ACTGGTTTCCCATTCATCTACACCCTGAAGGCCCCAACTGTCTTCAAATTCGAGGTTGAT 3685
1161 T  G  F  P  F  I  Y  T  L  K  A  P  T  V  F  K  F  E  V  D

     GCTTCTGCCAAGACCCACCCGCAAGTGTACAAGATGCCAGCTGGTCACCCAGAAACCGAA 3745
1181 A  S  A  K  T  H  P  Q  V  Y  K  M  P  A  G  H  P  E  T  E

     AACGACGATTTCTTCTATATGCCACAGTCTATTAACGGATCCGTCGATGTGAACCTGTTG 3805
1201 N  D  D  F  F  Y  M  P  Q  S  I  N  G  S  V  D  V  N  L  L

     TACCACCGCATGGTTGATGCCAAGGTTGGATTTGTCACTCCATTCGATCACCAACGTTAC 3865
1221 Y  H  R  M  V  D  A  K  V  G  F  V  T  P  F  D  H  Q  R  Y

     ATCGCTGGTTACCAGAAGAAGCTGCACGGTTATTTGCCCTTCAACGTTGAGTTGGGACTG 3925
1241 I  A  G  Y  Q  K  K  L  H  G  Y  L  P  F  N  V  E  L  G  L

     GACTTTGTCAAGGATGAGTATGAGTTCGAATTCAAATTCCTGGAACCCAAGGACGACCAT 3985
1261 D  F  V  K  D  E  Y  E  F  E  F  K  F  L  E  P  K  D  D  H

     CTGCTGTTCCACATGAGCTCGTGGCCATACACTGGATACAAGGACATTACCGATATGCGC 4045
1281 L  L  F  H  M  S  S  W  P  Y  T  G  Y  K  D  I  T  D  M  R

     CCGATTGCCGAAAATCCAAATGCCAAGATTGTGCACGATGACAACCAATCTACCAAGACC 4105
1301 P  I  A  E  N  P  N  A  K  I  V  H  D  D  N  Q  S  T  K  T

     ATGGAACACACTTTCGGTCAGGATATGACCGGTGTTGCTCTGCGATTCCACGCCAAATAC 4165
1321 M  E  H  T  F  G  Q  D  M  T  G  V  A  L  R  F  H  A  K  Y

     GACTTTGATCTGATCAACTTCCAACAGTTCTGGAGCTTGATTCAGAAGAACGACTTCGTT 4225
1341 D  F  D  L  I  N  F  Q  Q  F  W  S  L  I  Q  K  N  D  F  V

     TCGGCTGTGAACTATCCATTCGCTTACCAGCCATATGAATACCATCAGTTCAACCTGTTC 4285
1361 S  A  V  N  Y  P  F  A  Y  Q  P  Y  E  Y  H  Q  F  N  L  F

     TACGATTCCCAGCGTACTCACGCCAAATCGTTCAAGTTCTTCGCTTACCAGAAGTTCGGT 4345
1381 Y  D  S  Q  R  T  H  A  K  S  F  K  F  F  A  Y  Q  K  F  G

     GCTCCTTCTTTCGAAGAAACTGGACCGAAGCACCCAGCCAACCGTCACTCTTACTCTGGA 4405
1401 A  P  S  F  E  E  T  G  P  K  H  P  A  N  R  H  S  Y  S  G

     AACTATTACGAATCGAACTACGCTCAACCCTTCGTCTACAGCCCTGGAAGCCAACGCCGC 4465
1421 N  Y  Y  E  S  N  Y  A  Q  P  F  V  Y  S  P  G  S  Q  R  R

     TATGAACAATTCTTCCGCAATGCTGCTTCTGGAATCAGAAATAGCTTCGTTCGTTACTAC 4525
1441 Y  E  Q  F  F  R  N  A  A  S  G  I  R  N  S  F  V  R  Y  Y
```

Fig. 2 (continued)—*legend on p. 948.*

```
     GACTTCGGCTTCGAATTCTACGCTCCACAGTATAAGAGTGAATTTACTTTCACAACCGCT 4585
1461 D  F  G  F  E  F  Y  A  P  Q  Y  K  S  E  F  T  F  T  T  A

     TTCGCTGATAGTCCAGTTGACAAGACTTCCCGCCAGCTGTACTACTTCTATGCCAGCCCA 4645
1481 F  A  D  S  P  V  D  K  T  S  R  Q  L  Y  Y  F  Y  A  S  P

     ATGTTCCCAAGCCAATCGTACTTCAAGGATATTCCATTCAGTGGAAAGCAATTCCAGTTC 4705
1501 M  F  P  S  Q  S  Y  F  K  D  I  P  F  S  G  K  Q  F  Q  F

     TGCGCCACCGCTACCAGTGAATTCCCACGCGTTCCTTACCTGAAGTTCTCGGACTTTGAC 4765
1521 C  A  T  A  T  S  E  F  P  R  V  P  Y  L  K  F  S  D  F  D

     AAATACTACGGAGATGCTAGCCAGTACTTCGATTTCCTGTATGGTGAATCTTGCCAAGGA 4825
1541 K  Y  Y  G  D  A  S  Q  Y  F  D  F  L  Y  G  E  S  C  Q  G

     GGAGCTCACATTGCTGTGAAGGGTAAGCAGAAGCAGACTGGAAAGTGCCGCGAATACCTG 4885
1561 G  A  H  I  A  V  K  G  K  Q  K  Q  T  G  K  C  R  E  Y  L

     CGATTCTCGGATGTTGCTAAGGCTTGCAAGGAACAGATGGCCAACGGATACTACCAATTC 4945
1581 R  F  S  D  V  A  K  A  C  K  E  Q  M  A  N  G  Y  Y  Q  F

     GAGGAATGCCAACAGGCTATCGATCAGGCGTACTATTACGACTTCTACGATTACGCCATT 5005
1601 E  E  C  Q  Q  A  I  D  Q  A  Y  Y  Y  D  F  Y  D  Y  A  I

     GAGTACAAGGATGTCGGCTCTGTTGCCAAGAATCTGACCAACAAGTTCTACAACTACTTC 5065
1621 E  Y  K  D  V  G  S  V  A  K  N  L  T  N  K  F  Y  N  Y  F

     CAGTACGCGTTCTACCCGTACTTCGAATCGAACTTCTTCTACCATGGAAAGTCCAACTAC 5125
1641 Q  Y  A  F  Y  P  Y  F  E  S  N  F  F  Y  H  G  K  S  N  Y

     ATCAAGGCTGAATTCGAATTCGCTCCTTATGGCGATTACTACAACGCTTCGTTCTTCGGA 5185
1661 I  K  A  E  F  E  F  A  P  Y  G  D  Y  Y  N  A  S  F  F  G

     CCAAGCTACGCCTTCCAGGTTCAGAACTACCCGGTCTTCAATGACTACTCTACATACTTC 5245
1681 P  S  Y  A  F  Q  V  Q  N  Y  P  V  F  N  D  Y  S  T  Y  F

     CCATACTTCTTCAAGTACACTTTCTTCCCACGTTATCAACCATACTACATGCACCGTCTG 5305
1701 P  Y  F  F  K  Y  T  F  F  P  R  Y  Q  P  Y  Y  M  H  R  L

     CCATCGCACAAGCCCCGCAACCGTCCGTACTACGAGCTTTCCAACTACGAACAGTTCGCC 5365
1721 P  S  H  K  P  R  N  R  P  Y  Y  E  L  S  N  Y  E  Q  F  A

     ATCTTCGATCGTAAACCACAGTATCgtaagttcaagtaattctgtaagttttcaatttat 5425
1741 I  F  D  R  K  P  Q  Y  P

     taacgctaattgctttttttcagCTTCATGCTCTTTCTCCAACGACAACTTCTACACCTTC 5485
1750                        S  C  S  F  S  N  D  N  F  Y  T  F

     GACAACAAGAAGTACTTCTACGATATGGGAGAATGCTGGCATGCAGTGATGTACACTGTG 5545
1762 D  N  K  K  Y  F  Y  D  M  G  E  C  W  H  A  V  M  Y  T  V

     AAGCCAGACTACGACTTCTATGCTCAACAATCCCACTTCTACAACTCTGACTTTGAGTAC 5605
1782 K  P  D  Y  D  F  Y  A  Q  Q  S  H  F  Y  N  S  D  F  E  Y

     AAGTACAAGAATGGATTTGAAGAGTACGAACAGTTCGCTGCCCTGGCCCGTCGTGGATCT 5665
1802 K  Y  K  N  G  F  E  E  Y  E  Q  F  A  A  L  A  R  R  G  S

     GACAATCAGCTATACTTCAAGTTCTTGTTCGGAGACAACTACATTGAAGTTTTCCCCAAC 5725
```

Fig. 2 (continued)—*legend overleaf.*

```
1822 D  N  Q  L  Y  F  K  F  L  F  G  D  N  Y  I  E  V  F  P  N

     AACGGTGGTGTTCCATTTGTGAAGTACAACGGACGTCCATACGACATCAGCAAGAGCAAC 5785
1842 N  G  G  V  P  F  V  K  Y  N  G  R  P  Y  D  I  S  K  S  N

     ATTGCCCACTTCGAATACAAGGAAGGCTACCCAAGCTTCCCCTTCTTCTACGCTTTTGCC 5845
1862 I  A  H  F  E  Y  K  E  G  Y  P  S  F  P  F  F  Y  A  F  A

     TACCCCAACAAGGATTTGGAAGTCAGCTTCTTCGGTGGAAAACTGAAGTTCGCTACCGAT 5905
1882 Y  P  N  K  D  L  E  V  S  F  F  G  G  K  L  K  F  A  T  D

     GGATACCGTGCTCGCTTCTTCTCGGACTACTCGTTCTACAACAACTTTGTCGGTCTGTGC 5965
1902 G  Y  R  A  R  F  F  S  D  Y  S  F  Y  N  N  F  V  G  L  C

     GGAACCAACAACGGAGAATACTTCGATGAATTTGTCACCGCCGATCAGTGCTACATGCGC 6025
1922 G  T  N  N  G  E  Y  F  D  E  F  V  T  A  D  Q  C  Y  M  R

     AAACCTGAGTTCTTCGCTGCTTCCTACGCCATCACTGGACAGAACTGTACCGGACCAGCC 6085
1942 K  P  E  F  F  A  A  S  Y  A  I  T  G  Q  N  C  T  G  P  A

     AAGGCCTTCAACTATGCCTACCAACAGAAGGCCAAGCAGGAATGTGTCAAGCGTGAAGTC 6145
1962 K  A  F  N  Y  A  Y  Q  Q  K  A  K  Q  E  C  V  K  R  E  V

     TACTATGGAGACATCATCTACAACCAGGAATACTACCACCCCCGCTACCGCTACTACAAC 6205
1982 Y  Y  G  D  I  I  Y  N  Q  E  Y  Y  H  P  R  Y  R  Y  Y  N

     CACAATGTTGAAGAGTCCTCCAGCTCTTCGTCTAGCTCCTCTTCCGATTCTTCGTCCTCG 6265
2002 H  N  V  E  E  S  S  S  S  S  S  S  S  S  S  D  S  S  S  S

     TCATCTTCTTCGGAATTCAGCTCTCTGGGGCGCTCCGGCAGCTCATCGTCTAGCTCGAGC 6325
2022 S  S  S  S  E  F  S  S  L  G  R  S  G  S  S  S  S  S  S  S

     TCTGAGGAACAGAAGGAATTCCACCCACATAAGCAGGAACACAGCATGAAGGAATGCCCA 6385
2042 S  E  E  Q  K  E  F  H  P  H  K  Q  E  H  S  M  K  E  C  P

     GTTCAGCATCAGCACCAATTCTTCGAGCAAGGTGACCGCATTTGCTTCTCCCTGCGTCCT 6445
2062 V  Q  H  Q  H  Q  F  F  E  Q  G  D  R  I  C  F  S  L  R  P

     CTGCCAGTGTGCCACTCCAAGTGCGCTGCTACGGAAAAGATCAGCAAATACTTCGATGTC 6505
2082 L  P  V  C  H  S  K  C  A  A  T  E  K  I  S  K  Y  F  D  V

     CACTGCTTCGAGAAGGATTCCACCCAGGCTAAGAAGTACAAATCCGAGATTGGCCGCGGC 6565
2102 H  C  F  E  K  D  S  T  Q  A  K  K  Y  K  S  E  I  G  R  G

     TACACTCCGGACTTCAAGAGCTTCGCCCCACACAAGACTTACAAGTTCAACTACCCGAAG 6625
2122 Y  T  P  D  F  K  S  F  A  P  H  K  T  Y  K  F  N  Y  P  K

     AGCTGTGTCTACAAGGCATACTAGgaaacgacatttgcagatcccatttttgtatgacga 6685
2142 S  C  V  Y  K  A  Y

     accaatgaactaacgaaataaattataaggcaatttttaaatatgtgttgtttgaattcc 6745

     aatttacgaattgagtcgac                                          6765
```

Fig. 2 (continued).

FIGURE 2. Nucleotide sequence of 8780 bp of the VgA1 clone. Underlined regions are, in order, the TATA box, the transcription start site, the first ATG, the stop codon, and the polyA addition signal. N-terminal sequences of both the small and large subunits of the protein are double underlined. The imperfect triple repeats of 9–10 amino acid residues are asterisked. The introns are shown in lower case letters. Nucleotides are numbered at the right, in reference to the transcription start site as +1. Amino acids are numbered at the left.

Computer analysis

DNA sequence data were assembled using PC Gene and subsequently analyzed using the MacVector sequence analysis software of International Biotechnologies Inc. and the "Sequence Analysis Software Package" version 7.1 from the Genetics Computer Group (Devereux *et al.*, 1984). GCG programs used included: Pileup for multiple sequence alignment, Gap for pairwise sequence comparisons, Blast for database searches, Fetch for retrieving sequence data from GenBank, Findpattern for finding selected patterns in a particular sequence, Compare and Dotplot for comparison of two sequences using dotmatrix.

RESULTS AND DISCUSSION

Nucleotide sequence and the structure of the vitellogenin gene VgA1 in Aedes aegypti

Four of the five vitellogenin (Vg) genes of the mosquito, *A. aegypti* have been cloned and mapped (Gemmill *et al.*, 1986; Hamblin *et al.*, 1987). One of these, VgA1, has been used to follow vitellogenin mRNA levels in physiological experiments (Gemmill *et al.*, 1986; Racioppi *et al.*, 1986) and is the subject of this investigation. Figure 1 shows the revised restriction map of the original clone containing this gene. Since Northern blots (Gemmill *et al.*, 1986) suggested that the vitellogenin coding region began near the 3′ end of the EcoRI B-fragment, it seemed likely that important regulatory elements of the gene would be in this fragment. Therefore, 8780 bp of the vitellogenin A1 genomic clone was sequenced, including the EcoRI B fragment (Fig. 2). An open reading frame was found to begin near the 3′ end of the EcoR I B fragment and extend almost to the 3′ end of the fragment.

Primer extension was used to identify the transcription initiation site. A 21 base oligonucleotide complementary to sequence immediately upstream of the predicted ATG (see below) was hydridized to total RNA extracted from blood-fed females and extended using AMV reverse transcriptase. The results (Fig. 3) indicate a start site beginning ATCACTT, 75 nucleotides upstream of the ATG. This sequence corresponds well to the consensus mRNA cap site derived from *D. melanogaster*, ATCA[G/T]T[C/T] (Hultmark *et al.*, 1986), differing only by one nucleotide. Sequence TCACT (+2 to +6) is similar to the arthropod initiator consensus TCAGT found in the vicinity of the initiation site (Cherbas and Cherbas, 1993). A potential TATA box is located 30 nucleotides upstream from the cap site. This sequence differs from the consensus G/(A)TATAAAA (Breathnach and Chambon, 1981) by one base and is appropriately spaced upstream of the cap site. The ATG at +76 is the first possible translation start site downstream of the cap. This sequence, AAGATGC, contains an adenine at position −3 that is most critical for efficient translation initiation (Kozak, 1989), although it lacks a purine at +4.

A 70 base intron interrupting the open reading frame in the 11th codon was predicted by consensus rules (Breathnach and Chambon, 1981). Its presence was confirmed by RNase protection (Fig. 4) in which a major pair of fragments define the intron end points at positions +107 and +176. The minor pair of fragments could be due to expression of another member of the gene family, or to a different allele.

By comparison with the cDNA sequence of *A. aegypti* vitellogenin (Chen *et al.*, 1994), a second, 57 bp, intron is predicted between positions +5391 and +5447. Both the 5′ and 3′ splice sites for this intron conform

FIGURE 3. Determination of the transcription start site using primer extension. RNA from whole blood-fed females was used as the template. A 21 base ^{32}P-labeled oligonucleotide immediately upstream of the ATG was used as primer. Lanes 1–4 contain a sequencing reaction of the region of interest. The arrow head indicates the position of the TATA box. Lanes 6 and 7 contain the primer extended sample. Lane 8 contains a 1 kb ladder.

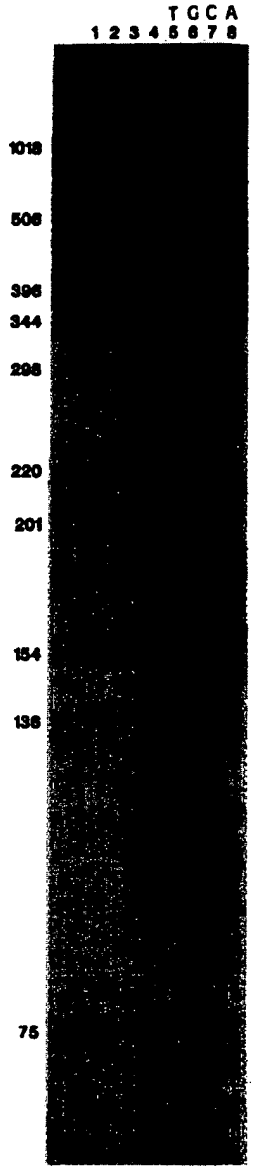


FIGURE 4. Location of an intron in the signal sequence using RNase protection. A ^{32}P-labeled riboprobe was synthesized using T3 RNA polymerase from the T3 promoter to the first Cla I site in the B fragment. The riboprobe was hybridized to RNA extracted from blood-fed females, digested with RNase A and RNase T1 prior to separation by PAGE. Lane 1, a 1 kb ladder. Lane 2, the probe. Lane 4, the RNase treated sample. Lanes 5–8 contain a sequencing reaction of the region of interest.

to the consensus rules of Breathnach and Chambon (1981).

A stop codon was found at +6647 followed by a poly A addition signal beginning at +6702. Therefore, the 8780 bp sequenced region of the VgA1 genome clone contained 2015 bp of 5′ untranscribed upstream sequence, a 6369 bp open reading frame interrupted by two introns, and a short 3′ untranslated region.

Protein product of the VgA1 gene

The deduced amino acid sequence is also shown in Fig. 2. Duplicated sequences were found, beginning at residue 29 which consisted of three sequential imperfect repeats:

$$_{29}\text{GYKGYDAGYK}_{38}$$
$$_{39}\text{GY GYDAGYK}_{47}$$
$$_{48}\text{GY GYDAGYK}_{56}$$

The nucleotide sequences of these repeats are slightly less similar than the amino acid sequence due to the degeneracy of codons.

It has been shown that the two subunits of the vitellogenin of *A. aegypti* originate from a common precursor (Bose and Raikhel, 1988). To precisely localize the beginning of the secreted peptides, we assumed that the N-terminus of vitellin would be the same as for vitellogenin. Vitellin was purified from mature eggs and the two subunits were separated by SDS–PAGE and electrophoretically blotted to PVDF membrane. N-terminal sequencing of protein subunits on the blots showed that the N-terminal sequence of the smaller subunit was: YQYENAFKGYNPGYK. As shown in Fig. 2, this sequence was found to begin at amino acid 17. There was a serine instead of an alanine, in the amino acid sqeuence deduced from DNA corresponding to position 6 of the above N-terminal sequence. This difference could arise because of different expressed alleles of the VgA1 gene, or the heterogeneity of vitellogenin due to expression of different members of the vitellogenin gene family. This evidence suggests that the region from the first methionine to amino acid 16 contains a signal sequence. This also conforms to the prediction of the −3, −1 rule of the signal peptide cleavage site (Von Heijne, 1990). The first intron interrupts the signal sequence in codon 11.

The N-terminal sequence of the large subunit was found to be: DLNAIKEKKYYEAYK. As shown in Fig. 2, this sequence is found to begin at amino acid residue 469. A cleavage sequence for dibasic processing endoproteases (Barr, 1991), RYRR, appears just before the beginning of the large subunit. This site was also located by Chen *et al.* (1994) using N-terminal sequences of vitellogenin, rather than vitellin, suggesting that the N-termini of the vitellogenin subunits are not modified after uptake into the egg.

Variations in the mosquito VgA1 gene

The mosquito vitellogenin cDNA has been sequenced by Chen *et al.* (1994). The cDNA clones isolated by Chen *et al.* were identified using the EcoRI C fragment of the VgA1 gene as a probe. The Rockefeller strain of *A. aegypti* was also used by this group (A. Raikhel, personal communication). They found six base substitutions near the 5′ end of their 6504 bp cDNA sequence, which were suggested to be due to allelic differences in the mosquito population. None of these base substitutions affected the amino acid residues. Their cDNA sequence also showed several single nucleotide differences from our genomic sequence at nucleotides +68, +2863, +4326, +4872, +5471, +6005, +6064,

+6292, +6293, +6294, +6295. All 11 of these substitutions are at different positions from their six substitutions. These 11 substitutions result in changes in six amino acid residues. As described above, three imperfect repeats of 10 amino acids were found in our sequence near the beginning of the small subunit from residues 29 to 56 while the cDNA sequence reported by Chen *et al.* (1994) showed only two of these imperfect repeats. The extra copy of the repeat is not likely to be an intron because it does not fit the highly conserved 5′ and 3′ intron splicing sequence, and 27 bp is too small to be an intron. Mount *et al.* (1992) showed that the smallest mRNA intron in *D. melanogaster* was 51 nucleotides by surveying 209 introns in the entire database. Our survey of all 18 introns of three mosquito genera (*Aedes*, *Anopheles*, and *Culex*) in the database showed that intron size in mosquitoes ranged from 52 to 1737, with a median of 65 nucleotides.

The only differences between the sequence of the coding region of our genomic clone and the sequence of the cDNA of Chen *et al.* (1994) are the above mentioned substitutions and duplication. Of the four members of the vitellogenin gene family, only VgA2 has a similar restriction map to VgA1 which is restricted to the EcoRI C-fragment (Hamblin *et al.*, 1987). It is very likely that their cDNA and our genomic sequence are from the same vitellogenin gene, VgA1. Therefore, in addition to the variation at the 5′ end of the cDNA described by Chen *et al.* (1994), more substitutions and a different number of duplications are present, suggesting that the VgA1 gene is polymorphic at many positions.

Conservation at insect vitellogenins

As described above, several genomic and cDNA sequences of insect vitellogenins have been reported. With these sequences, and the genomic sequence of *A. aegypti* vitellogenin available, gene organization as well as the deduced amino acid sequences of vitellogenins of these insects can be compared.

Conservation of the overall primary structure among insect vitellogenins. As shown in Fig. 5(A), the entire amino acid sequences of the three insect vitellogenins aligned with each other very well using dot matrix comparisons. The entire deduced amino acid sequence of mosquito vitellogenin showed 49.9% similarity (S) to the boll weevil vitellogenin and 50.6% to the silkworm vitellogenin. The boll weevil vitellogenin showed 48.1% similarity to the silkworm vitellogenin. The percentage of overall identity (I) of each pair of sequences are also shown in the legend of Fig. 5. The significance of the similarities of these sequences were tested by analyzing the quality of each comparison and the average quality of 100 randomized comparisons. The adjusted quality of each comparison was calculated using the following formula of Gribskov and Burgess (1986):

Q' (adjusted quality) = [Q (quality of each comparison)—A (average quality of 100 randomization)]/SD (standard deviation of the 100 randomization).

Generally, the Q' value is less influenced by sequence length. A Q' of 3.0 or higher is required for confidence that two sequences are significantly related (Gribskov and Burgess, 1986). Sequences that are not related have a low Q'. For example, *A. aegypti* vitellogenin VgA1 and *Homo sapiens* serum albumin have a Q' of −0.78. As shown in the legend of Fig. 5(A), all three comparisons of insect vitellogenins are highly significant.

As seen in the dot matrix, some regions in the vitellogenins are more similar than others. The region with the highest sequence identity was found near the N-terminal end of the large subunits of all three vitellogenins. Yano *et al.* (1994b) also described a conserved region in a similar position. As shown in Fig. 6, 27.0% of the amino acids were identical between all three sequences, and 69.8% were identical between at least two of them. There are also many conserved substitutions. The similarities between the three vitellogenins in this region were 64.4% (*A. aegypti* vs *A. grandis*), 60.7% (*A. aegypti* vs *B. mori*), and 57.5% (*A. grandis* vs *B. mori*), respectively. It is not known if this region has a functional role, however, it is likely that regions with specific functions such as binding to the vitellogenin receptor would be more conserved.

In many insects, vitellogenins are composed of two types of subunits, one with a molecular weight higher than 120 kDa, and one with a molecular weight around 55 kDa (Kunkel and Nordin, 1985). It has been shown that in the mosquito, sawfly, boll weevil and silkworm, that the small and large subunits of the vitellogenins are both encoded on a single mRNA. The four amino acid residues preceding the cleavage sites between the small and large subunits are relatively conserved among all four insects as shown in Fig. 7. The four residues conform to the consensus cleavage sequence for dibasic processing endoproteases, R(K)XXR (Barr, 1991). In the mosquito, sawfly, and silkworm, the sequences flanking the cleavage site contain polyserine tracts which are longer in the mosquito and sawfly than in the silkworm. The flanking sequence in the boll weevil is also relatively serine rich, however, no polyserine tract is present. In the mosquito the high serine content (9.9%) is mainly due to three polyserine regions, two of which flank the cleavage site between the subunits. The third one is close to the C-terminal end. The silkworm and the boll weevil vitellogenins contain 9.8 and 7.9% serine residues respectively. The partial sequence of the sawfly vitellogenin also showed a very high percentage of serines. Vertebrate vitellogenins are also rich in serine, however, this is due to a single long polyserine run in the region of phosvitin whose position does not match that of the insect polyserine regions.

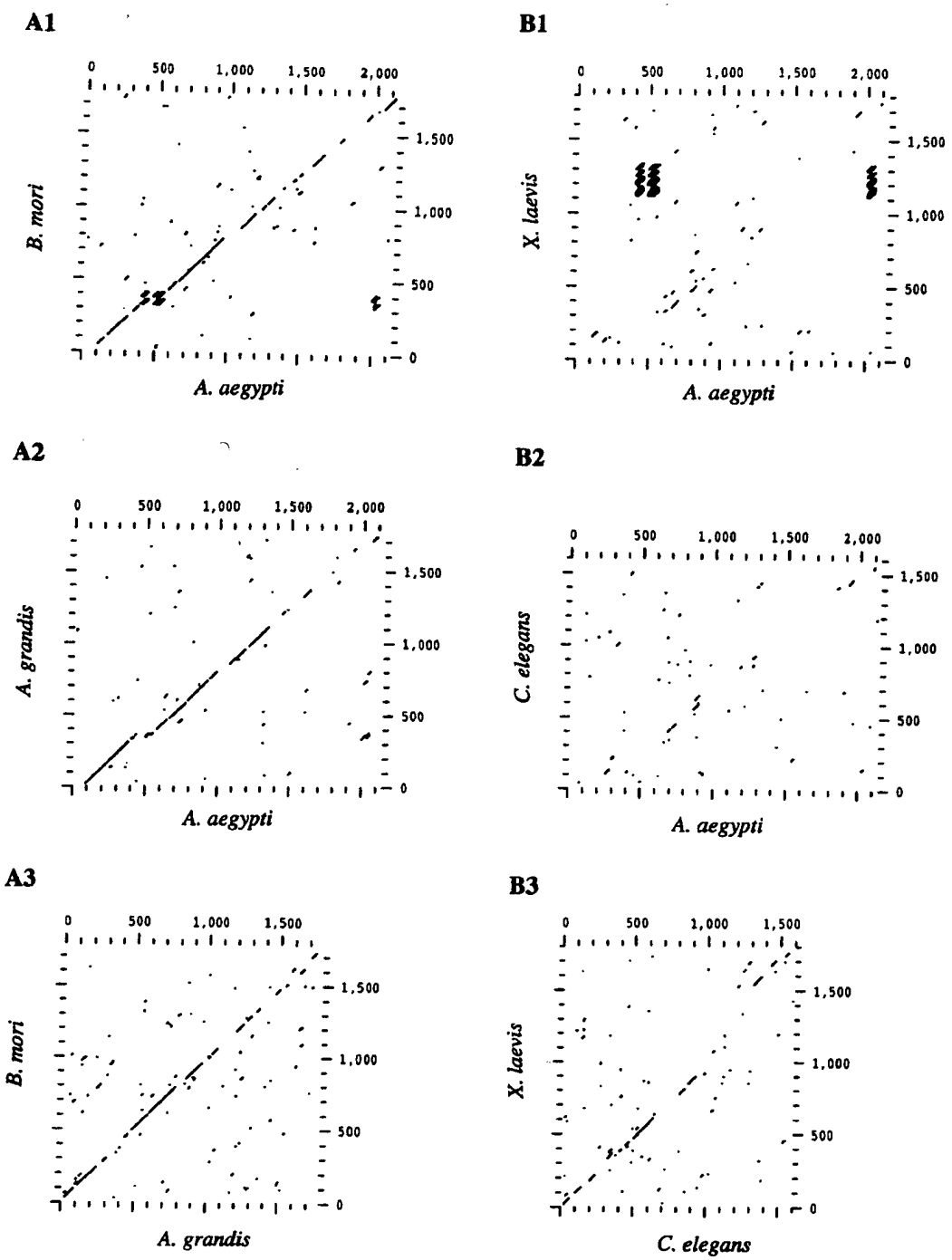


FIGURE 5. (A, 1–3) Pairwise dot matrix comparisons of the complete amino acid sequences of the vitellogenins of *A. aegypti*, *A. grandis*, and *B. mori*. (B, 1–3) Pairwise dot matrix comparisons of the complete amino acid sequences of the vitellogenins of *A. aegypti*, *X. laevis*, and *C. elegans*. These were done using Compare and Dotplot of GCG. Window size was 40, and the stringency was 19. Percentage similarity (S), percentage identity (I), quality of the comparison (Q), and average quality of 100 randomized comparisons (A), and the adjusted quality of the comparison (Q') as defined in the text were also calculated using GAP of GCG, shown as the following:

Figure	S	I	Q	A	Q'
A 1	50.6%	29.1%	935.9	527.8 ± 8.0	51
2	49.9	29.0	975.4	520.4 ± 8.3	55
3	48.1	26.5	875.3	519.2 ± 7.3	49
B 1	40.6	18.7	563.7	527.4 ± 8.0	4.5
2	44.6	21.0	553.0	482.5 ± 6.7	11
3	46.2	22.9	646.2	485.8 ± 7.0	23

```
A.aegypti        WVTLRDAV AEAGTPSAFK LIFDFIKEKK LRGYEAATVI ASLAQSIRYP TEHLLHEFFL
A.grandis        WSIFRDSV AEAGTGPALL NIKKWIETKK IQKTEAAQVI GTLAQSTRFP TEEYMRKFFE
B. mori          WMIFRDGV TQAGTLPAFK QIQSWIENKK IQEEEAAQVV VALPRTLRYP TKQIMTQFFN
CONSENSUS(2)     W-IFRD-V AEAGT-PAFK -I--WIE-KK IQ--EAAQVI --LAQS-RYP TE--M--FF-
CONSENSUS(3)     W---RD-V --AGT--A-- -I---I--KK ----EAA-V- --L---R-P T------FF-

LVTSDVVLHQ EYLNATALFA YSNFVNQAHV SNRSAYNYYP VFSFGRL.AD ADYKIIEHKI VPWFAHQLRE
LATETQVRQQ ETLNQTCILS YTNLVHKVYI NRNESHNQFP VHAFGSFYTK KGREFVKTTV IPHLKQELEK
FARSPAVKDQ MFLNSSALMA ATKLINLGQV NNYTAHSYYP THMYGRL.TH KHDAFVLEEI LPTLAADLKA
LATS--V--Q E-LN-TAL-A YTNLVN---V NN--AHNYYP VH-FGRL-T- K---FV---I -P-LA--L--
------V--Q --LN------ ---------- ---------P ----G----- ---------- -P-----L--

AVNEGDSVKI QVYIRSLGNL GHPQILSVFE PYLEGTIQIT DFQRLAIMVA LDNLVIYYPS LARSVLYRAY
AISNADNNKI HVMIRALGNI GHKSILNVFQ PYFEGEKQVS QFQRLMMVAC MDRLADCYPH IARSVFYKIY
TVEYKDSTKA QVYIQAIGNL GHREILKVFA PYLEGKVEIS TYLRTHIVKN LKTLAKLRDR HVRAVLFSIL
AV---DS-KI QVYIRALGNL GH--IL-VF- PYLEG--QIS -FQRL-IV-- LD-LA--YP- -ARSVLY-IY
-----D--K- -V-I---GN- GH--IL-VF- PY-EG----- ---R------ ---L------ --R-V-----

QNTADVHEVR CAAVHLLMRT DPPADMLQRM AEFTHHDPRL YVRAAVKSAI
QNTAELPEIR VVAVHQLIRA NPPVEMLQRM AQYTNTDSQE EVNAAVKSVI
RNTAEPYPVR VAAIQSIFIS HPTGEMMQAM AEMTHNDPSV EVRAVLKSAI
QNTAE--EVR VAAVH-L-R- -PP-EMLQRM AE-TH-DP-- EVRAAVKSAI
-NTA-----R --A------- -P---M-Q-M A--T--D--- -V-A--KS-I
```

FIGURE 6. Deduced amino acid sequence comparisons within the most conserved regions in the three insect vitellogenins. This region is located near the N-terminal of the large vitellogenin subunits. The first three rows are the amino acid sequences of vitellogenins of *A. aegypti*, *A. grandis* and *B. mori*, respectively. The fourth row is the sequence conserved between two insects (69.8% identity), and the last row is the sequence conserved in all three insects (27.0% identity).

Runs of polyserines often contain many phosphorylation sites. Chen *et al.* (1994) suggested that the clusters of serines in the mosoquito vitellogenin could help assimilate the high level of phosphate taken up with the blood meal. However, as described above, polyserine tracts are seen in the vitellogenins of insects with different feeding habits.

The locations of cysteines in vitellogenins of the nematode and the vertebrates are relatively conserved, especially at the C-termini of the proteins (Spieth *et al.*, 1991). These authors suggested that this may be due to a requirement to form particular tertiary structures dependent on the disulfide bonds. The numbers of cysteines in the vitellogenins of the mosquito, boll weevil, and the silkworm are 20, 20, and 14, respectively. Among these, seven are at the same positions in all three insects and an additional seven are at the same positions in two insects. All seven cysteine positions that are conserved in the three insects, and five out of the seven that are conserved in two insects, are concentrated near the C-terminus within approx. 550 amino acid residues.

Conservation of the signal peptides and positions of introns. Shown in Fig. 8 is a comparison of signal sequences of seven vitellogenins with six yolk proteins of the Cyclorraphid Diptera. It is clear that the signal sequences are similar within the sets of vitellogenins and yolk proteins, but not between them. Some of the residues are conserved, particularly the leucines at positions 6, 7, 9, 10 and 13 (Fig. 8). Within the seven

```
A. aegypti   YVRPGNETSS SSSSSSSSSS SSSESSSSSS ESVENPKISP VEQYKPLLDK VEKRGNRYRR
A. grandis   NNQQQQPEEL SNPLDIGNLV YTYGQPKNNQ VHSKLNENLM EDSSSEESSE QEMTHRRFRR
A. rosae           No sequence available<--S SSSSSSSEED QQRGNSENKN NNQHGHRQAR
B. mori      AEWPRAGAMR PAQSILYSLS TKQMTKHYES SSSSSSSESH EFNFPEQHEH PHQSNQRGRR

DLNAIKEKKY YEAYKMDQYR LHRLNDTSSD SSSSDSSSSS SSESKEHRNG TSSYSSSSSS SSSSSSSESS
SANSLTKQWR ESSEEWNQQQ QQPRPQLTRA PHSPLLPSMV GYHGKSIKEN KDFDIRQNVE NLVTEISDEI
AANQRGQGNQ DSDSSSSSSS SSSSSSSSDS SSSSSSESDS LSSSEEYWQS RPTLTDAPEA PMLPLFIGYK
SYMRSKLVTV HKVLKKRNSE SSSGSSSSSA DSSSTYINDD IPDIDEPAYA ALYMSPQPHA DKKQNAMNAQ

SYSSSSSSSS ESYSISSEEY
KQSEKTISKH TLDKYTILNT
GSAAQQSSQV NPVSVAKKLA
KILQDIAQQL QNPNNMPKSD
```

FIGURE 7. Deduced amino acid sequence comparison of the regions near the cleavage site between the small and large vitellogenin subunits of four species of insects showing serine rich regions flanking the cleavage site. The four amino acid residues preceding the cleavage site (in bold) are aligned. Serines are underlined.

A. aegypti	VgA1	M L A K L L L L A L A G L T A A
*A. gambiae**	Vg 1&2	M I A K L L L L T L V G L C T
*L. migratoria**	A	M W A V I V L G L L intron
	B	M W A L I L S G L L intron
A. grandii		M W S T V A L C L L V G L S Y V S S
A. rosae **		M W S P L L L C L L V G I A S A
B. mori		M K L F V L A A I I A A V S S
Conserved residues in insect vitellogenins		W A L L L L V G L
*D. melanogaster**	YP 1	M N P M R V L S L L A C L A V A A L A
	YP 2	M N P L R T L C V M A C L L A V A M G
	YP 3	M M S L R I C L L A T C L L V A A H A
*C. capitata**	Vg 1	M N P L K I F C F L A L V I A V A S A
	Vg 2	M N P L T I F C L V A V L L S A A T A
C.erythrocephala	YP b	M N P L R I V C V A L L L A A A G S A
Conserved residues in Cyclorraphid Diptera yolk proteins		N P L R I C A L A A A
Conserved residues in vertebrate vitellogenins		R G I I L A L L L A L A G S

FIGURE 8. The signal sequence of the *A. aegypti* VgA1 precursor protein compared with signal sequences from other insects. Conserved residues among either the seven insect vitellogenins or the six Cyclorraphid Diptera yolk proteins, defined as being present in more than 50% of the sequences, are shown below. The conserved residues in vertebrate vitellogenin are also shown (Blumenthal and Zuker-Aprison, 1987). An underlined amino acid indicates the position of an intron within that codon. The asterisk indicates that the exact end of the signal sequence has not been determined. The double asterisk indicates sequences derived from a cDNA, thus information about intron position is not available. The *Anopheles gambiae* sequence is from P. Romans (unpublished observations). Other signal sequences are from: boll weevil, *A. grandis* (Trewitt *et al.*, 1992); silkworm, *B. mori* (Yano *et al.*, 1994a); locust, *L. migratoria* (Locke *et al.*, 1987); saw fly, *A. rosae* (Kageyama *et al.*, 1994); the fruit fly *D. melanogaster* (Hung *et al.*, 1983; Garabedian *et al.*, 1987), the Mediterranean fruit fly, *C. capitata* (Rina and Savakis, 1991) and the blowfly, *C. erythrocephala* (Martinez and Bownes, 1994). The length of the intron in the locust genes has not been determined.

vitellogenin signal sequences, six contain an intron in the same location, 10 residues from the initial methionine. The size of this intron varies considerably from 70 in *A. aegypti* to 2057 in *A. grandis*. The residue at position 11 (at the intron insertion) is either an alanine or a valine. The lengths of the signal sequences of these vitellogenins are between 15 and 18 amino acids.

The signal sequence of yolk protein b (YPb) of the *C. erythrocephala* is 19 amino acids long (Martinez and Bownes, 1994). The lengths of the signal sequences of the other five yolk proteins are unknown because the N-terminal sequences of the mature proteins have not been reported. As shown in Fig. 8, these six signal sequences are very similar, however, the conserved residues are not similar to those in the vitellogenins of other insects, except for the leucine at position 13 and alanine at position 11.

Signal sequences of vertebrate vitellogenin genes are highly conserved (Spieth *et al.*, 1985; Blumenthal and Zuker-Aprison, 1987). Most of the conserved residues are different from those conserved in the insect vitellogenins and those in yolk proteins. However, some of the residues conserved in vertebrates are also conserved in the seven insect vitellogenins, particularly the leucines at positions 6, 9 and 10 (Fig. 8).

The mosquito VgA1 gene has two introns. There are six introns in both boll weevil and silkworm vitellogenin genes. As described above, the position of intron 1 in the signal sequence is conserved. The position of the second intron in the mosquito is conserved with intron 6 of the boll weevil, and intron 5 of the silkworm, splitting the codon of a proline. The position of intron 3 of the boll weevil is also conserved with intron 2 of silkworm, splitting the codon of a leucine. There are, therefore, three conserved positions of introns in these insects. Trewitt *et al.* (1992) found that the second and third conserved positions in the boll weevil are also conserved in intron 12 and 29 of the chicken and frog vitellogenin genes. However, because the similarities between insect and vertebrate vitellogenins are limited, the comparison of intron position between insects and vertebrates may be imprecise.

Relationship between insect vitellogenins and vitellogenins of vertebrates and other invertebrates, and the yolk proteins of Cyclorraphid Diptera

As described above, insect vitellogenins are very similar in their primary structures, suggesting that they came from a common ancestor. There is also limited similarity between insect, nematode and vertebrate vitellogenins. Figure 5(B) shows the dot matrix comparisons of vitellogenins between *A. aegypti* and *Xenopus laevis*, between *A. aegypti* and *Caenorhabditis elegans*, and between *X. laevis* and *C. elegans*. The percentage similarity and

identity of each comparison, and the quality and the adjusted quality of each comparison are shown in the figure legend. Clearly, the nematode vitellogenins are more similar to those of vertebrates than insects. However, Q' values for the comparisons of *A. aegypti* vs *X. laevis*, and *A. aegypti* vs *C. elegans* are both higher than 3.0, suggesting that they are also related.

It has been suggested that the yolk proteins in Cyclorraphid Dipteran insects such as *D. melanogaster* (Hung *et al.*, 1983; Garabedian *et al.*, 1987), *C. capitata* (Rina and Savakis, 1991) and *C. erythrocephala* (Martinez and Bownes, 1994) are related to mammalian triacylglycerol lipase (Baker, 1988; Bownes *et al.*, 1988; Terpstra and AB, 1988). It has been shown that yolk proteins in these insects are very closely related to each other in terms of amino acid sequence (Rina and Savakis, 1991; Martinez and Bownes, 1994). Our analysis indicates that similarities between these yolk proteins and vitellogenins of other insects are limited. For example, all three yolk proteins of *D. melanogaster* matched to a region around the cleavage site of *A. aegypti* vitellogenin, with 38–40% similarity, with 12–19 gaps inserted. The similarities are relatively low considering the length of sequences compared. The Q' values of these comparisons are 0.47 (for YP 1), 0.57 (YP 2), 1.73 (YP 3). All are less than 3.0, suggesting that the similarities between these pairs are not significant. Therefore, the relationship of the yolk proteins of Cyclorraphid Diptera and the vitellogenins of other insects is very distant.

The phylogenetic relationship between vitellogenins of insects, other invertebrates, vertebrates, and the yolk proteins of the Cyclorraphid Diptera is a rather complex question. Simple sequence comparisons are not sufficient to address these problems: detailed phylogenetic approaches will need to be employed. Meanwhile, the accumulation of molecular data in this field will be very useful for this analysis.

Analysis for potential regulatory regions in the VgA1 gene

Several lines of evidence suggest that the expression of vitellogenin genes of *A. aegypti* is under the control of 20-hydroxyecdysone (ecdysone) (Spielman *et al.*, 1971; Hagedorn *et al.*, 1973, 1975; Bohm *et al.*, 1978; Borovsky *et al.*, 1985; Gemmill *et al.*, 1986; Racioppi *et al.*, 1986; Ma *et al.*, 1987). Several regulatory elements have been identified in steroid hormone controlled genes. Ecdysone receptors of *D. melanogaster* and *Chironomus tentans* have been cloned and sequenced (Koelle, 1991; Imhof *et al.*, 1993). The ecdysone receptor in *D. melanogaster* functions as a heterodimer with ultraspiracle (USP), a member of the retinoid X receptor family of transcription factors (Yao *et al.*, 1992; Thomas *et al.*, 1993; Antoniewski *et al.*, 1994). Antoniewski *et al.* (1993) proposed a revised consensus ecdysone response element (EcRE) [PuG(G/T)T(C/G)A(N)T G(C/A)(C/A)(C/t)Py] for *D. melanogaster* based on mutational experiments. EcR and ultraspiracle (USP) heterodimer binding sites (EcR/USP) have also been identified (Antoniewski *et al.*, 1994). Moreover, three relatively long regions that control tissue specific expression (fat body enhancers) have been identified in the YP1, YP2, and Fbp 1 genes of *D. melanogaster* (Garabedian *et al.*, 1986; Abrahamsen *et al.*, 1993; Antoniewski, 1994). A 9 bp consensus sequence was found in the regulatory regions of YP1–3 genes of *D. melanogaster* (Logan and Wensink, 1990). Other transcription factors such as the CCAAT/enhancer binding protein (C/EBP), and doublesex proteins (DSX) have also been shown to bind to the regulatory regions of these yolk protein genes (Burtis *et al.*, 1991; Falb and Maniatis, 1992). C/EBP is involved in the acquisition of responsiveness to steroid hormones in some genes in mammals (Ben-Or and Okret, 1993). DSX proteins have been shown to regulate female specificity of the expression of the yolk proteins in *D. melanogaster* (Burtis *et al.*, 1991). Furthermore, Fos and Jun binding proteins (AP-1 family of transcription activators) that are known to interact with glucocorticoid receptors (Miner and Yamamoto, 1991) in yeast and mammals, have also been found in *D. melanogaster* (Perkins *et al.*, 1988). Finally, hormone response elements (HREs) for many other members of the steroid hormone receptor superfamily, such as glucocorticoid receptor (GR), estrogen receptor (ER), thyroid hormone receptor (TR), vitamin D receptor (VDR), retinoid X receptor (RXR), retinoic acid receptor (RAR), have also been identified (Martinez and Wahli, 1991).

Using computer programs including Findpattern, Bestfit, and Gap of the GCG software package, we searched the vitellogenin gene for evidence of potential regulatory elements, focusing on the 2015 bp of 5′ upstream region, the introns, and the 3′ untranslated region. Relatively stringent criteria were used and sites that did not show the correct spacing were eliminated. Two regions in the EcoR I B-fragment were found with multiple copies of sequences with high degrees of identity to the consensus sequences of different hormone response elements. These elements are more numerous in region 1 (−76 to −457 bp) than in region 2 (−708 to −1106). Most of the elements in region 1 also have higher degree of identity to the consensus sequences than those in region 2. Moreover, the fat body enhancers of *D. melanogaster* YP1 (127 bp, Garabedian *et al.*, 1986) and YP2 (343 bp, Abrahamsen *et al.*, 1993) are similar to the sequence around region 1, although the similarity is not statistically significant.

Shown in Fig. 9 is the overall structure of the EcoR I B-fragment and a detailed illustration of the putative regulatory regions 1 and 2. Within region 1, three short stretches of sequences were found to contain several HREs, shown as A, B, C. In addition to the HREs shown in A, B, C, three other HREs including a copy of C/EBP binding site (−457 to −449) were also found within region 1. Within region 2, two short stretches of sequences were found to contain multiple HREs, shown as D, E. In addition to the HREs shown in D and E, three copies of EcR/USP binding sites, one DSX binding

site, and one ERE were also found in other areas of region 2.

An EcRE (+105 to +117, 11 of 13 bp are identical to the consensus) was found near the 5′ end of the first intron. Moreover, a C/EBP site (+103 to +111, 9 of 9 identical) overlaps this EcRE. An EcRE (+5394 to +5406, 11 of 13 identical), and a C/EBP site (+5405 to +5413, 9 of 9 identical) were also present in the second intron. It is also noteworthy that two DSX sites (+253 to 261, 7 bp of 9 identical; and +280 to +288, 8 bp of 9 identical) were found in the coding region where the three imperfect repeats are located. Several HREs were also found in the 116 bp 3′ untranslated region including two EcREs (+6684 to +6696 and +6731 to +6743, both 11/13), one EcR/USP site (+6730 to +6744, 11 of 15 identical), two Fos binding sites (+6680 to +6686, and +6757 to 6763, both 6 of 7 identical) and two CSDs (+6684 to +6692, 8 of 9 identical, and +6754 to +6762, 7 of 9 identical). Deitsch and Raikhel (1993) found four repeated putative regulatory elements with variable levels of similarity to retinoic acid response elements (RAREs) in the vitellogenic carboxypeptidase gene, a fat body specific, female specific, ecdysone controlled gene in *A. aegypti*. A single RARE (10/12) was found in the VgA1 gene (see Fig. 9) with the correct spacing between half sites.

The presence of these conserved sites may be fortuitous and must be confirmed with functional evidence, which is currently being sought.

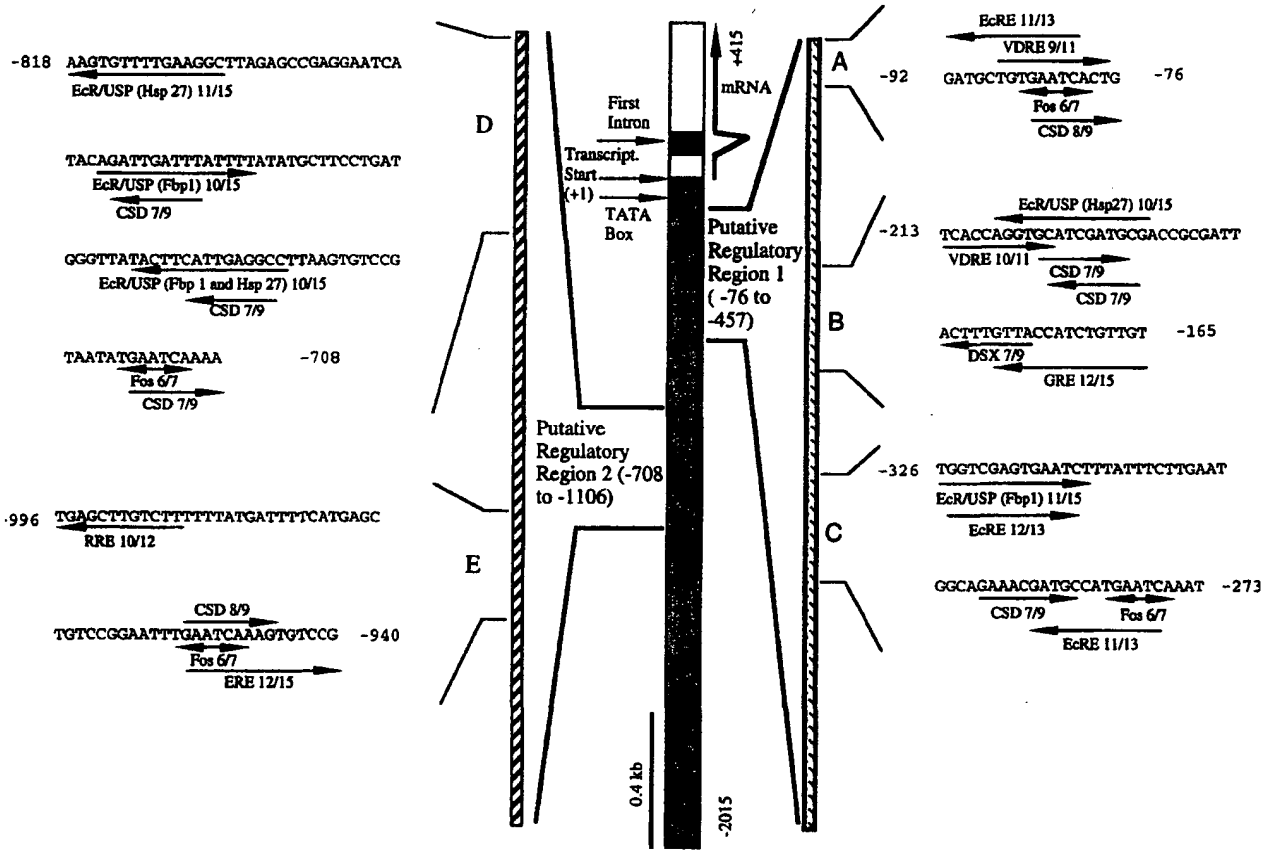


FIGURE 9. Potential regulatory sequences in the EcoR I B-fragment of the VgA1 gene of *A. aegypti*. Putative regulatory region 1 is shown with three short stretches of sequences (A, B, C) where regulatory elements concentrate. Putative regulatory region 2 is shown with two short stretches of sequences (D, E) where regulatory elements concentrate. Arrows under the sequence indicates the locations where specific hormone response elements match. The direction of the arrows represents the direction of the binding site. CSD: consensus sequence of the enhancer regions of *D. melanogaster* yolk proteins 1-3, GAATCAATG (Logan and Wensink, 1990); DSX: consensus binding site for double sex binding protein, CTACAAAGT (Burtis *et al.*, 1991); EcRE: Ecdysone response elements, RG[G/T]T[C/G]ANTG[C/A][C/A][C/t]Y (Antoniewski *et al.*, 1993); EcR/USP (FBp1): EcR/USP heterodimer binding site identified in the *D. melanogaster* Fbp1 gene, GGGTTGAATGAATTT (Antoniewski *et al.*, 1994); EcR/USP (Hsp 27): EcR/USP heterodimer binding site identified in the *D. melanogaster* Hsp 27 gene, GGGTTCAATGCACTT (Antoniewski *et al.*, 1994); ERE: Estrogen response element, N_A GGTCAN_CNN$_G$TGACCN_T (Martinez and Wahli, 1991); Fos: binding site for Fos protein, TGACTCA; GRE: glucocorticoid response element, AG[A/G]ACANNNTGTACC (Martinez and Wahli, 1991); RRE [RARE]: retinoic acid response element, [A/G]GG[T/A]CA in direct repeats or in palindromes (Martinez and Wahli, 1991); VDRE: vitamin D response element, GGTGANTCACC, or TCACCNGGTGA (Martinez and Wahli, 1991). Other consensus sequences mentioned in the text, which are not shown in Fig. 9 include, C/EBP binding site of *D. melanogaster* YP1 gene, TGTTGCAAT (Falb and Maniatis, 1992); and C/EBP binding site consensus in mammals, T[T/G]NNG[C/T]AA[T/G] (Falb and Maniatis, 1992).

REFERENCES

Abrahamsen N., Martinez A., Kjær T., Sondergaard L. and Bownes M. (1993) Cis-regulatory sequences leading to female-specific expression of yolk protein genes 1 and 2 in the fat body of *Drosophila melanogaster*. *Molec. Gen. Genet.* **237**, 41–48.

Antoniewski C., Laval M. and Lepesant J.-A. (1993) Structural features critical to the activity of an ecdysone receptor binding site. *Insect Biochem. Molec. Biol.* **23**, 105–114.

Antoniewski C., Laval M., Dahan A. and Lepesant J.-A. (1994) The ecdysone response enhancer of the Fbp1 gene of *Drosophila melanogaster* is a direct target for the EcR/USP nuclear receptor. *Molec. Cell. Biol.* **14**, 4465–4474.

Baker M. E. (1988) Is vitellogenin an ancestor of apolipoprotein B-100 of human low-density lipoprotein and human lipoprotein lipase? *Biochem. J.* **255**, 1057–1060.

Barr P. J. (1991) The long-sought dibasic processing endoproteases. *Cell* **66**, 1–3.

Ben-Or S. and Okret S. (1993) Involvement of a C/EBP-like protein in the acquisition of responsiveness to glucocorticoid hormones during chick neural retina development. *Molec. Cell. Biol.* **13**, 331–340.

Blumenthal T. and Zuker-Aprison E. (1987) Evolution and regulation of vitellogenin genes. In *Molecular Biology of Invertebrate Development*, pp. 3–19. Alan R. Liss, New York.

Bohm M. K., Behan M. and Hagedorn H. H. (1978) Termination of vitellogenin synthesis by mosquito fat body, a programmed response to ecdysterone. *Physiol. Ent.* **3**, 17–25.

Borovsky D., Thomas B. R., Carlson D. A., Whisenton L. R. and Fuchs M. S. (1985) Juvenile hormone and 20-hydroxyecdysone as primary and secondary stimuli of vitellogenesis in *Aedes aegypti*. *Archs Insect Biochem. Physiol.* **2**, 75–90.

Bose S. G. and Raikhel A. S. (1988). Mosquito vitellogenin subunits originate from a common precursor. *Biochem. Biophys. Res. Commun.* **155**, 436–442.

Bownes M., Shirras A., Blair M., Collins J. and Coulson A. (1988) Evidence that insect embryogenesis is regulated by ecdysteroids released from yolk proteins. *Proc. Natn. Acad. Sci.* **85**, 1554–1557.

Breathnach R. and Chambon P. (1981) Organization and expression of eukaryotic split genes coding for proteins. *A. Rev. Biochem.* **50**, 349–383.

Burtis K. C., Coschigano K. T., Baker B. S. and Wensink P. C. (1991) The doublesex proteins of *Drosophila melanogaster* bind directly to a sex-specific yolk protein gene enhancer. *EMBO J.* **10**, 2577–2582.

Chen J-S., Cho W.-L. and Raikhel A. S. (1994) Analysis of mosquito vitellogenin cDNA, similarity with vertebrate phosvitins and arthropod serum proteins. *J. Molec. Biol.* **237**, 641–647.

Cherbas L. and Cherbas P. (1993) The arthropod initiator: the capsite consensus plays an important role in transcription. *Insect Biochem. Molec. Biol.* **23**, 81–90.

Chomczynski P. and Sacchi N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analyt. Biochem.* **162**, 156–159.

Corthesy B., Leonnard P. and Wahli W. (1990) Transcriptional potentiation of the vitellogenin B1 promoter by a combination of both nucleosome assembly and transcription factors: an *in vitro* dissection. *Molec. Cell. Biol.* **10**, 3926–3933.

Deitsch K. W. and Raikhel A. S. (1993) Cloning and analysis of the locus for mosquito vitellogenic carboxypeptidase. *Insect Molec. Biol.* **2**, 205–213.

Devereux J., Haeberli P. and Smithies O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucl. Acids Res.* **12**, 387–395.

Falb D. and Maniatis T. (1992) A conserved regulatory unit implicated in tissue-specific gene expression in *Drosophila* and man. *Genes Dev.* **6**, 454–465.

Garabedian M. J., Shepherd B. M. and Wensink P. C. (1986) A tissue-specific enhancer from the *Drosophila* yolk protein 1 gene. *Cell* **45**, 859–867.

Garabedian M. J., Shirras A. D., Bownes M. and Wensink P. C. (1987) The nucleotide sequence of the gene coding for *Drosophila melanogaster* yolk protein 3. *Gene* **55**, 1–8.

Gemmill R. M., Hamblin M., Glaser R. L., Racioppi J. V., Marx J. L., White B. N., Calvo J. M., Wolfner M. F. and Hagedorn H. H. (1986) Isolation of mosquito vitellogenin genes and induction of expression by 20-hydroxyecdysone. *Insect Biochem.* **16**, 761–774.

Gribskov M. and Burgess R. (1986) Sigma factors from *E. coli*, *B. subtilis*, phage SP01, and phage T4 are homologous proteins. *Nucl. Acids Res.* **14**, 6745–6763.

Hagedorn H. H. (1985) The role of ecdysteroids in reproduction. In *Comprehensive Insect Physiology, Biochemistry and Pharmacology* (Edited by Kerkut G. A. and Gilbert L. I.), Vol. 8, pp. 205–262. Pergamon Press, Oxford.

Hagedorn H. H. (1994) The endocrinology of the adult female mosquito. In *Advances in Disease Vector Research*, Vol. 10, pp. 109–148. Springer, New York.

Hagedorn H. H. and Fallon A. M. (1973) Ovarian control of vitellogenin synthesis by the fat body in *Aedes aegypti*. *Nature* **244**, 103–105.

Hagedorn H. H., O'Connor J. D., Fuchs M. S., Sage B., Schlaeger D. A. and Bohm M. K. (1975) The ovary as a source of alpha ecdysone in an adult mosquito. *Proc. Natn. Acad. Sci. U.S.A.* **72**, 3255–3259.

Hamblin M. T., Marx J. L., Wolfner M. F. and Hagedorn H. H. (1987) The vitellogenin gene family of *Aedes aegypti*. *Mem. Inst. Oswaldo Cruz* **82**, Suppl. III, 109–114.

Hiremath S., Lehtoma K. and Nagarajan M. (1994) cDNA cloning and expression of the large subunit of juvenile hormone-regulated vitellogenin from *Lymantria dispar*. *J. Insect Physiol.* **40**, 813–821.

Hultmark D., Klemenz R. and Gehring W. H. (1986) Translational and transcriptional control elements in the untranslated leader of the heat-shock gene *hsp 22*. *Cell* **44**, 429–438.

Hung M.-C. and Wensink P. C. (1983) Sequence and structure conservation in yolk proteins and their genes. *J. Molec. Biol.* **164**, 481–492.

Imhof M. O., Rusconi S. and Lezzi M. (1993) Cloning of a *Chironomus tentans* cDNA encoding a protein (cEcRH) homologous to the *Drosophila melanogaster* ecdysteroid receptor (dEcR). *Insect Biochem. Molec. Biol.* **23**, 115–124.

Kageyama Y., Kinoshita T., Umesono Y., Hatakeyama M. and Oishi K. (1994) Cloning of cDNA for vitellogenin of *Athalia rosae* (Hymenoptera) and characterization of the vitellogenin gene expression. *Insect Biochem. Molec. Biol.* **24**, 599–605.

Koelle M. R., Talbot W. S., Segraves W. A., Bender M. T., Cherbas P. and Hogness D. S. (1991) The *Drosophila* EcR gene encodes an ecdysone receptor, a new member of the steroid receptor superfamily. *Cell* **67**, 59–77.

Koeppe J. K., Fuchs M., Chen T. T., Hunt L.-M., Kovalick G. E. and Briers T. (1985) The role of juvenile hormone in reproduction. In *Comprehensive Insect Biochemistry and Pharmacology* (Edited by Kerkut G. A. and Gilbert L. I.), Vol. 8, pp. 165–203.

Kozak M. (1989) The scanning model for translation: an update. *J. Cell Biol.* **108**, 229–241.

Kunkel J. G. and Nordin J. H. (1985) Yolk proteins. In *Comprehensive Insect Physiology Biochemistry and Pharmacology* (Edited by Kerkut G. A. and Gilbert L. I.), Vol. 1, pp. 83–111. Pergamon Press, New York.

Locke J., White B. N. and Wyatt G. R. (1987) Cloning and 5′ end nucleotide sequences of two juvenile hormone-inducible vitellogenin genes of the African migratory locust. *DNA* **6**, 331–342.

Logan S. K. and Wensink P. C. (1990) Ovarian follicle cell enhancers from the *Drosophila* yolk protein genes: different segments of one enhancer have different cell-type specificities that interact to give normal expression. *Genes Dev.* **4**, 613 (1990).

Ma M., He G., Zhang J. Z. and Gwadz R. (1987) Response of cultured *Aedes aegypti* fat bodies to 20-hydroxyecdysone. *J. Insect Physiol.* **33**, 89–93.

Martinez A. and Bownes M. (1994) The sequence and expression pattern of the *Calliphora* erythrocephala yolk protein A and B genes. *J. Molec. Evol.* **38**, 336–351.

Martinez T. and Hagedorn H. H. (1987) Development of responsiveness to hormones after a blood meal in the mosquito, *Aedes aegypti*. *Insect Biochem.* **17**, 1095–1098.

Martinez E. and Wahli W. (1991) Characterization of hormone response elements. In *Nuclear Hormone Receptors* (Edited by Parker M. G.) pp. 125–153. Academic Press, London.

Miner J. N. and Yamamoto K. R. (1991) Regulatory cross talk at composite response elements. *Trends Biochem. Sci.* **16**, 423–426.

Mount S. M., Burks C., Hertz G., Stormo G. D., White O. and Fields C. (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucl. Acids Res.* **20**, 4255–4262.

Nardelli D., van het Schip F. D., Gerber-Huber S., Haefliger J. A., Gruber M., AB G. and Whali W. (1987) Comparison of the organization and fine structure of a chicken and a *Xenopus laevis* vitellogenin gene. *J. Biol. Chem.* **262**, 15,377–15,385.

Perkins K. K., Dailey G. M. and Tjian R. (1988) Novel Jun- and Fos-related proteins in *Drosophila* are functionally homologous to enhancer factor AP-1. *EMBO J.* **7**, 4265–4273.

Racioppi J. V., Gemmill R. M., Kogan P. H., Calvo J. M. and Hagedorn H. H. (1986) Expression and regulation of vitellogenin messenger RNA in the mosquito, *Aedes aegypti*. *Insect Biochem.* **16**, 255–262.

Rina M. and Savakis C. (1991) A cluster of vitellogenin genes in the Mediterranean fruit fly, *Ceratitis capitata*: sequence and structural conservation in Dipteran yolk proteins and their genes. *Genetics* **127**, 769–780.

Shapiro J. P. and Hagedorn H. H. (1982) Juvenile hormone and the development of ovarian responsiveness to brain hormone in the mosquito *Aedes aegypti*. *Gen. Comp. Endocrinol.* **46**, 176–183.

Spielman A., Gwadz R. W. and Anderson W. A. (1971) Ecdysone-initiated ovarian development in mosquitoes. *J. Insect Physiol.* **17**, 1807–1814.

Spieth J., Denison K., Kirtland S., Cane J. and Blumenthal T. (1985) The *C. elegans* vitellogenin genes: short sequence repeats in the promotor regions and homology to the vertebrate genes. *Nucl. Acids Res.* **13**, 5283–5295.

Spieth J., Nettleton M., Zucker-Aprison E., Lea K. and Blumenthal T. (1991) Vitellogenin motifs conserved in nematodes and vertebrates. *J. Molec. Evol.* **32**, 429–438.

Terpstra P. and AB G. (1988) Homology of *Drosophila* yolk proteins and the triacylglycerol lipase family. *J. Molec. Biol.* **202**, 663–665.

Thomas H. E., Stunneberg H. G. and Stewart A. F. (1993) Heterodimerization of the *Drosophila* ecdysone receptor with retinoid X receptor and *ultraspiracle*. *Nature* **362**, 471–475.

Trewitt P. M., Heilmann L. J., Degrugillier S. S. and Kumaran A. K. (1992) The boll weevil vitellogenin gene: nucleotoide sequence, structure, and evolutionary relationship to nematode and vertebrate vitellogenin genes. *J. Molec. Evol.* **34**, 478–492.

Von Heijne G. (1990) The signal peptide. *J. Membr. Biol.* **115**, 195–202.

Yano K.-I., Sakurai M. T., Watabe S., Izumi S. and Tomino S. (1994a) Structure and expression of mRNA for vitellogenin in *Bombyx mori*. *Biochim. Biophys. Acta* **1218**, 1–10.

Yano K.-I., Sakurai M. T., Watabe S., Izumi S. and Tomino S. (1994b) Vitellogenin gene of the silkworm, *Bombyx mori*: structure and sex-dependent expression. *FEBS Lett.* **356**, 207–211.

Yao T.-P., Segraves W. A., Oro A. E., McKeown M. and Evans R. M. (1992) *Drosophila* ultraspiracle modulates ecdysone receptor function via heterodimer formation. *Cell* **71**, 63–72.

Acknowledgements—We thank Beipi Chen for preparing the vitellin for N-terminal sequencing and Michael Wang for preparation of sequencing template DNA and computer assembly of the sequence. This work was supported by NIH grants (No. HD 24869) to HHH and A. M. Fallon and Natural Sciences and Engineering Research Council and Insect Biotech Canada grants to PR.

The sex-determining gene *doublesex* in the fly *Megaselia scalaris*: Conserved structure and sex-specific splicing

Sylvia Kuhn, Volker Sievert, and Walther Traut

Abstract: The well-known sex-determining cascade of *Drosophila melanogaster* serves as a paradigm for the pathway to sexual development in insects. But the primary sex-determining signal and the subsequent step, *Sex-lethal* (*Sxl*), have been shown not to be functionally conserved in non-*Drosophila* flies. We isolated *doublesex* (*dsx*), which is a downstream step in the cascade, from the phorid fly *Megaselia scalaris*, which is a distant relative of *D. melanogaster*. Conserved properties, e.g., sex-specific splicing, structure of the female-specific 3′ splice site, a splicing enhancer region with binding motifs for the TRA2/RBP1/TRA complex that activates female-specific splicing in *Drosophila*, and conserved domains for DNA-binding and oligomerization in the putative DSX protein, indicate functional conservation of *dsx* in *M. scalaris*. Hence, the *dsx* step of the sex-determining pathway appears to be conserved among flies and probably in an even wider group of insects, as the analysis of a published cDNA from the silkmoth indicates.

Key words: sex-determining cascade, splice regulation, DNA-binding domain, oligomerization.

Résumé : La cascade génétique qui détermine le sexe chez le *Drosophila melanogaster* est bien connue et elle sert de paradigme pour l'étude du déterminisme sexuel chez les insectes. Cependant, il a été montré que le signal primaire du déterminisme sexuel et l'étape subséquente, *Sex-lethal* (*Sxl*), ne sont pas conservés chez les mouches n'appartenant pas au genre *Drosophila*. Les auteurs ont isolé le gène *doublesex* (*dsx*), lequel intervient en aval dans la cascade, chez la mouche phoridé *Megaselia scalaris*, laquelle est un lointain cousin du *D. melanogaster*. L'homologue putatif *dsx* montre des propriétés conservées (l'épissage spécifique en fonction du sexe, la structure du site d'épissage en 3′ spécifique aux femelles, une région d'accroissement de l'épissage comprenant des motifs de liaison au complexe TRA2/RBP1/TRA qui active l'épissage spécifique aux femelles chez la drosophile) et la protéine putative comprend des domaines conservés pour la liaison à l'ADN et pour l'oligomérisation. Tout cela suggère une conservation de la fonction de *dsx* chez le *M. scalaris*. Ainsi, l'étape *dsx* de la cascade semble être conservée chez les mouches et vraisemblablement chez un nombre encore plus grand d'insectes comme le suggère l'analyse d'une séquence d'ADNc du vers à soie.

Mots clés : cascade du déterminisme sexuel, régulation de l'épissage, domaine de liaison à l'ADN, oligomérisation.

[Traduit par la Rédaction]

Introduction

Sex-determining mechanisms defy the expectation that common basic biological functions use common genetic pathways. Primary sex-determining signals, as well as subsequent signal processing steps can be different even in related species. Flies are a group in which various primary sex-determining mechanisms have been discovered. In *Drosophila melanogaster*, the primary sex-determining signal is the ratio of X chromosomes to autosome sets, the X/A ratio (Bridges 1925). A number of X-chromosomal so-called numerator and one or more autosomal denominator genes are interacting to decide upon the female or male developmental pathway of the embryo (review: Cline and Meyer 1996). In *Chrysomya rufifacies*, sex is determined by a maternal factor; unisexual progenies, all-female or all-male, are produced depending solely on the genotype of the mother (Ullerich 1984). Presence or absence of an epistatic *Maleness* factor is the primary sex-determining signal in several other fly species, e.g. *Ceratitis capitata* (Willhoeft and Franz 1996), *Lucilia cuprina* (Bedo and Foster 1985), or *Megaselia scalaris* (Mainx 1966). In *M. scalaris*, the *Maleness* factor can change its location in a transposon-like fashion within the genome (Traut and Willhoeft 1990; Traut 1994). In the house fly, *Musca domestica*, several primary sex-determining mechanisms coexist (Dübendorfer et al. 1992).

Another known sex-determining cascade is that of the nematode *Caenorhabditis elegans*. Though resembling that of *D. melanogaster* superficially, transmission of the primary signal is mediated via steps of transcriptional regulation instead of splice regulation, and by a non-homologous set of

Corresponding Editor: P.B. Moens.

Received July 25, 2000. Accepted August 29, 2000. Published on the NRC Research Press web site November 7, 2000.

S. Kuhn, V. Sievert,[1] and W. Traut.[2] Institut für Biologie, Medizinische Universität Lübeck, Ratzeburger Allee 160, D-23538 Lübeck, Germany.

[1]Present address: MPI für Molekulare Genetik, Ihnestraße 73, D-14195 Berlin, Germany.

[2]Author to whom all correspondence should be addressed (e-mail: traut@molbio.mu-luebeck.de).

genes (reviewed in Kuwabara 1999). Comparative analyses in other fly species indicate that the sex-determining pathway of *Drosophila* is not even conserved among flies. Apart from the primary signal, the subsequent step in the cascade is also different. *Sex-lethal* (*Sxl*) serves this function and is sex-specifically spliced in *Drosophila* but not in non-*Drosophila* species like *Chrysomya rufifacies* (Müller-Holtkamp 1995), *Ceratitis capitata* (Saccone et al. 1998), *Musca domestica* (Meise et al. 1998), and *Megaselia scalaris* (Sievert et al. 1997; Sievert et al. 2000). Hence, it does not transmit the sex-determining signal. In contrast, *doublesex* (*dsx*), which functions as a double switch at the bottom of the cascade, initiating either female or male somatic development, appears to be structurally and functionally conserved in the Queensland fruit fly *Bactrocera tryoni* (Shearman and Frommer 1998) and *M. scalaris* (Sievert et al. 1997, this paper).

Here, we report on the functional organization of *dsx* in *M. scalaris* and discuss the role of *dsx* in sex determination of insects.

Material and methods

General methods

Genomic DNA and RNA were isolated from adult flies of the *M. scalaris* wild-type strain Wien, which has been kept in the laboratory for more than 30 years (Mainx 1964). Culture conditions were as described by Willhoeft and Traut (1990).

For molecular methods, we used standard procedures (Sambrook et al. 1989) unless otherwise noted. Genomic DNA from female and male flies was isolated as described by Blin and Stafford (1976). Total RNA was isolated with the Trizol reagent (Life Technologies, Eggenstein, Germany). Poly(A)$^+$ RNA was prepared from total RNA using the PolyATtract mRNA Isolation System IV (Promega Biotech, Madison, Wis.). The following oligonucleotides were custom-synthesized by MWG Biotech (Ebersberg, Germany): DSXf562R (TTCCTGGGAGGATCGTAGGG), DSXFbr (CTATTCCATTCCGGTTTCCG), DSXgb996R (GTTGCAAAACATAGAGTGGC), DSXK362R (ATCAAATTTCATATGGCAAG), and DSXK788F (TTGAGCAATTTCGATTTCCC).

For automated sequencing, we used the ABI PRISM™ Sequenase Terminator Double-Stranded DNA Sequencing Kit and the ABI PRISM™ dRhodamine Terminator Cycle Sequencing Ready Reaction Kit (PE Applied Biosystems, Weiterstadt, Germany).

Northern and Southern hybridization

Poly(A)$^+$ RNA from adult female and male *M. scalaris* was separated in denaturing agarose gels and blotted to a non-charged nylon membrane (Schleicher and Schüll, Dassel, Germany) using a vacuum blotting device and 20× SSPE (3.6 M NaCl, 200 mM NaH_2PO_4, 20 mM EDTA, pH 7.4) as transfer buffer. Hybridization was performed at 42°C for 24 h in a solution containing 45% (v/v) deionized formamide, 5× SSC [1× SSC is 150 mM NaCl, 15 mM sodium citrate, pH 7], 5× Denhardt's solution [1× Denhardt's solution is 0.1% polyvinylpyrrolidone (w/v), 0.1% (w/v) BSA, 0.1% (w/v) Ficoll 400], 0.1% (w/v) SDS, and 100 µg/mL salmon testis DNA (sonicated and denatured; Sigma-Aldrich Chemie GmbH, Deisenhofen, Germany). Washes were performed at 42°C in 0.1× SSC, 0.1% (w/v) SDS.

For Southern blots, 5 µg genomic DNA from female or male flies, was digested with *Bam*HI, *Bgl*II, *Eco*RI, *Hin*dIII, or *Xba*I, and separated in a 1% agarose gel. After 40 min incubation in a denaturing buffer containing 0.5 M NaOH and 1.5 M NaCl, the DNA was blotted to a non-charged nylon membrane using a vacuum device and the denaturing buffer as the transfer solution. Hybridization was performed at 68°C for 24 h in a solution containing 0.5 M Na_2HPO_4–NaH_2PO_4 (pH 7.2) and 7% (w/v) SDS. Washes were performed at 68°C in 0.1 M Na_2HPO_4–NaH_2PO_4 (pH 7.2), 0.1% (w/v) SDS.

Probe DSX5CS is a *Hin*dIII fragment of pMSWc178 (position 1–773) labeled by random priming with [α-^{32}P]dCTP. Southern hybridization was visualized using a Bio-Imaging Analyzer BAS-1000 (Fujifilm) and the accompanying analytical software PCBAS (Raytest, Straubenhardt, Germany).

Isolation of a genomic sequence

A genomic fragment of *M. scalaris dsx* was obtained by nested PCR on 300 ng genomic DNA from female and male flies with primers DSXK788F/DSXf562R in the first and DSXK788F/DSXFbr in the second round of amplification. We used 5 U Taq DNA polymerase in the buffer provided by the supplier (Life Technologies, Eggenstein, Germany), and 200 µM of each dNTP, 1.5 mM $MgCl_2$, and 0.2 µM of each primer. Cycling parameters in the first round were: 4 min initial denaturation at 94°C, hot start, 30 cycles of 30 s at 94°C, 30 s at 58°C, and 3 min at 72°C. In the second round: 4 min initial denaturation at 94°C, hot start, 30 cycles of 30 s at 94°C, 30 s at 55°C, and 1.5 min at 72°C, followed by 10 min final extension at 72°C. Four independent clones with identical intron sequence were isolated: pMSW2471 and pMSW2472 from DNA of female flies, pMSW2473 (Acc. No. AF283697) and pMSW2474 from DNA of male flies.

Reverse transcription PCR (RT–PCR)

Poly(A)$^+$ RNA was reverse transcribed using a *Not*I(dT)$_{18}$ primer and the First Strand cDNA Synthesis Kit (Pharmacia AB, Uppsala, Sweden). PCR was performed with 5 U Taq DNA polymerase in the buffer provided by the supplier, 200 µM of each dNTP, 1.5 mM $MgCl_2$, and 0.2 µM of each primer. The cycling parameters for PCR were: 4 min initial denaturation at 94°C followed by 38 cycles with 45 s at 94°C, 45 s at 53°C (primers DSXgb996/DSXK362R) or 60°C (DSXK788F/DSXf562R), 1.5 min at 72°C, and a final extension of 10 min at 72°C.

Construction of a cDNA library and screening for *M. scalaris dsx*

For construction of cDNA libraries from female and male *M. scalaris*, we used the Gigapack III Gold Cloning Kit (Stratagene, La Jolla, Calif.) according to the manufacturer's instructions. Roughly 1×10^6 plaques were screened with a RACE (rapid amplification of cDNA ends) generated cDNA probe from *M. scalaris dsx* (Sievert et al. 1997).

Sequence tools

BLASTX 2.0.11 (Altschul et al. 1997) was used for searches against the nr database (containing all nonredundant GenBank CDS translations, PDB, SwissProt, PIR and PRF entries), TBLASTN 2.0.14 (Altschul et al. 1997) for searches against the nr database and the dbEST database at NCBI (Bethesda, Md., accessed July 2000, http://www.ncbi.nlm.nih.gov/BLAST/ Altschul et al. 1997). Multiple sequence alignment was carried out using CLUSTALW (http://www2.ebi.ac.uk/clustalw, Thompson et al. 1994) using default parameters, the BLOSUM62 matrix (Henikoff and Henikoff 1992), and subsequent manual improvement. Phylogenetic analysis of protein sequences was performed with the neighbor-joining method of Saitou and Nei (1987), provided by the CLUSTALW tool. The resulting tree was visualized with TREEVIEW (Page 1996, http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

Fig. 1. (A) Structure of *Megaselia scalaris dsx* cDNA. pMSWc178 was isolated from the cDNA library of female flies, pMSWc233 from the library of male flies. Triangles, start and stop codons of the ORFs; horizontal arrows, primer positions; vertical arrows, polyadenylation signals. (B) Composition of the putative female-specific protein DSX^f and the male-specific protein DSX^m.

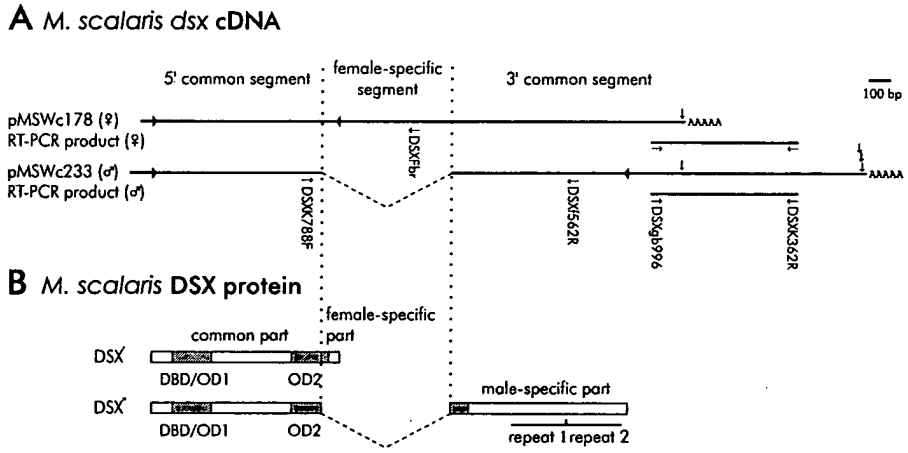


Fig. 2. RT–PCR on poly(A)$^+$ RNA from female and male flies (A) with primers DSXgb996 defined from the 3′ common segment and DSXK362R from the 3′ extension of pMSWc233 (B) with primers DSXK788F from the 5′ common segment and DSXf562R from the 3′ common segment of *Megaselia scalaris dsx*. + and – indicate presence or absence of reverse transcriptase in the reaction mix.

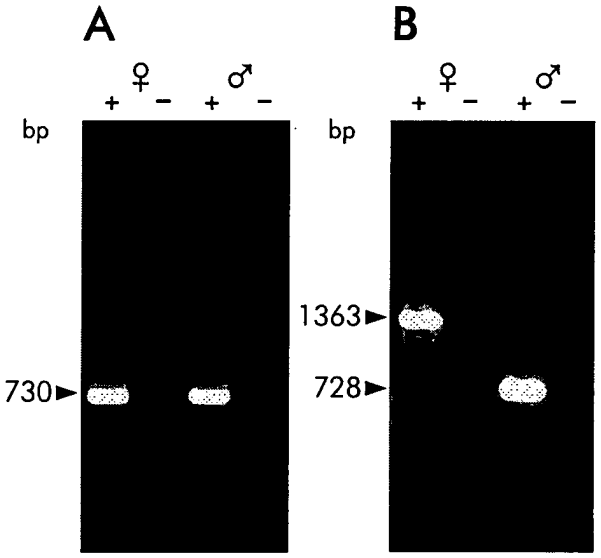
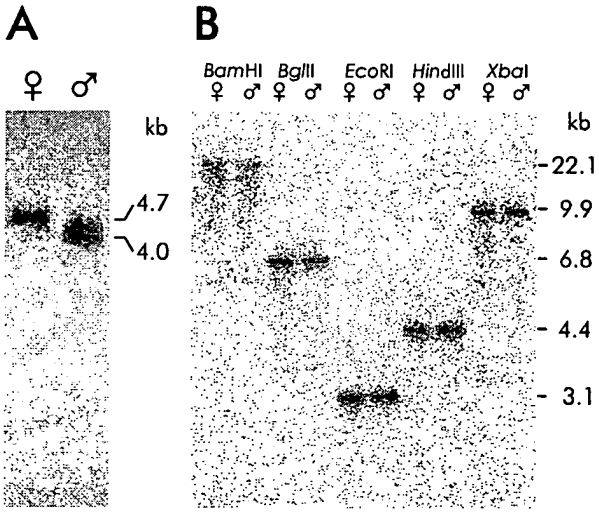


Fig. 3. (A) Northern blot of poly(A)$^+$ RNA from female and male flies, hybridized with probe DSX5CS from the 5′ common segment of *Megaselia scalaris dsx*; autoradiography. (B) Southern hybridization of DSX5CS to genomic DNA from female and male flies digested with restriction enzymes as indicated; phospho-imaging.



Results

RNA and cDNA of *M. scalaris dsx*

A 189-bp cDNA fragment of *M. scalaris dsx* had been isolated previously (Sievert et al. 1997). Using 3′ RACE on RNA of female flies with primers defined from this fragment yielded a cDNA of approximately 780 bp (Sievert et al. 1997). This was used as a probe to screen cDNA libraries from poly(A)$^+$ RNA of female and male *M. scalaris*. One clone of 2.7 kb (pMSWc178, Acc. No. AF283695) was isolated from a cDNA library of female flies and three clones of 2.4 kb (pMSWc184), 2.5 kb (pMSWc234), and 3.0 kb (pMSWc233, Acc. No. AF283696) from a cDNA library of male flies. The three cDNA clones from males were identical, except for the different extensions towards the 5′ end. All four clones contained the previously isolated 189-bp fragment of *M. scalaris dsx*.

Similarly to *dsx* in *D. melanogaster* and *B. tryoni* (Burtis and Baker 1989; Shearman and Frommer 1998), *M. scalaris dsx* transcripts are different in females and males. Three segments can be distinguished in *dsx* cDNAs of *M. scalaris* (Fig. 1A). The *dsx* cDNAs of female and male flies share a 5′ common segment, which in the longest cDNA has a length of 949 bp. They differ in a female-specific segment of 635 bp (position 899–1533 in pMSWc178) which follows the 5′

Fig. 4. Alignment of the putative DSX protein sequences from *Megaselia scalaris* (Acc. No. AF283695), *Drosophila melanogaster* (Acc. No. P23022), *Bactrocera tryoni* (Acc. No. AF029675), and *Bombyx mori* (Acc. No. AV398350). (A) The part of the proteins that is common to DSX^f and DSX^m; (B) female-specific part of DSX^f, the *B. mori* sequence is truncated; (C) male-specific part of DSX^m from *M. scalaris* (Acc. No. AF283696), *D. melanogaster* (Acc. No. P23023), and *B. tryoni* (Acc. No. AF029676). Black boxes, amino acids identical to the *M. scalaris* sequence; shaded boxes, amino acids similar to the *M. scalaris* sequence; DBD/OD1, DNA-binding and oligomerization domain 1; asterisks, residues whose replacement abolishes DNA binding activity; OD2, oligomerization domain 2 (the extension of the male-specific part of OD2 in *M. scalaris* is not clear, the label refers to *D. melanogaster* and *B. tryoni*); arrowheads, additional conserved region.

common segment in the female-derived cDNA, but is absent in the three male-derived clones. The remaining 3′ common segment up to the polyadenylation site in the female-derived clone is present in all cDNAs. The three male-derived cDNAs are extended 904 bp downstream of that point, but this 3′ extension is neither part of the open reading frame (ORF) nor male specific.

RT–PCR with primers DSXgb996/DSXK362R (for primer positions, see Fig. 1A) amplified a fragment of the 3′ extension in poly(A)$^+$ RNA from both female and male flies (Fig. 2A; Fig. 1A, RT–PCR product). Presence or absence of the 3′ extension in the 3′ common segment appears to depend on the use of two alternative (but not sex-specific) polyadenylation sites: the first one with a polyadenylation signal 20 bp before the poly(A) stretch in pMSWc178, the second with a series of three polyadenylation signals at 25 bp, 29 bp, and 39 bp upstream of the poly(A) tail in pMSWc233. In spite of the multiple polyadenylation signal, poly(A) tails start at identical positions in the three male-derived cDNA clones.

We performed RT–PCR on poly(A)$^+$ RNA from females and males with primers DSXK788F/DSXf562R, which spanned the female-specific segment. It amplified fragments compatible with the expected sizes of 1363 bp from females and 728 bp from males (Fig. 2B). This is evidence for the presence of the female-specific segment in female and its absence in male poly(A)$^+$ RNA.

Northern hybridization with probe DSX5CS from the 5′ common segment detected one prominent transcript in poly(A)$^+$ RNA from each of the female and male flies (Fig. 3A). The female-specific transcript, *dsx*^{f}, was roughly 0.7 kb larger than the male-specific transcript, *dsx*^{m}. The size difference corresponds to the size of the female-specific segment of 635 bp and confirms the presence of sex-specific *dsx* transcripts in females and males.

The *dsx* poly(A)$^+$ RNAs are approximately 1 kb larger than the cDNAs plus a putative poly(A) tail. Hence, although the cDNA clones contain the complete ORF (see below), they are incomplete at the 5′ untranslated region (UTR). Probably, the difference is accounted for by a long 5′ UTR, similar to that in *D. melanogaster* (Burtis and Baker 1989).

Megaselia scalaris dsx is a single-copy gene

Probe DSX5CS from the 5′ common segment was hybridized to genomic DNA of female and male flies, digested with one of five different restriction enzymes that do not cut within the probe sequence (Fig. 3B). The probe detected one fragment in each DNA sample. Therefore, *M. scalaris dsx* is most likely a single-copy gene, and the sex-specific transcripts *dsx*^{f} and *dsx*^{m} are alternative splice products.

Table 1. Six sequences from the female-specific exon of *Megaselia scalaris dsx* with similarity to the 13-nt splicing enhancer repeat elements of *Drosophila melanogaster dsx* (Burtis and Baker 1989; Inoue et al. 1992).

Position	Sequence	Identity
247	TCTTCAATCAACA	13/13
291	cCATCAATCAACA	12/13
320	TCATaAgTCAACA	11/13
275	TCAACAtTCAAtc	10/13
602	atATCAATCAAtA	10/13
269	caTTCA-TCAACA	10/13

Note: Position, distance downstream of the alternative splice site of the female-specific exon; identity, number of nucleotides (upper case letters) matching the consensus TCWWCRATCAACA (where W is A or T, and R is A or G) from *D. melanogaster*.

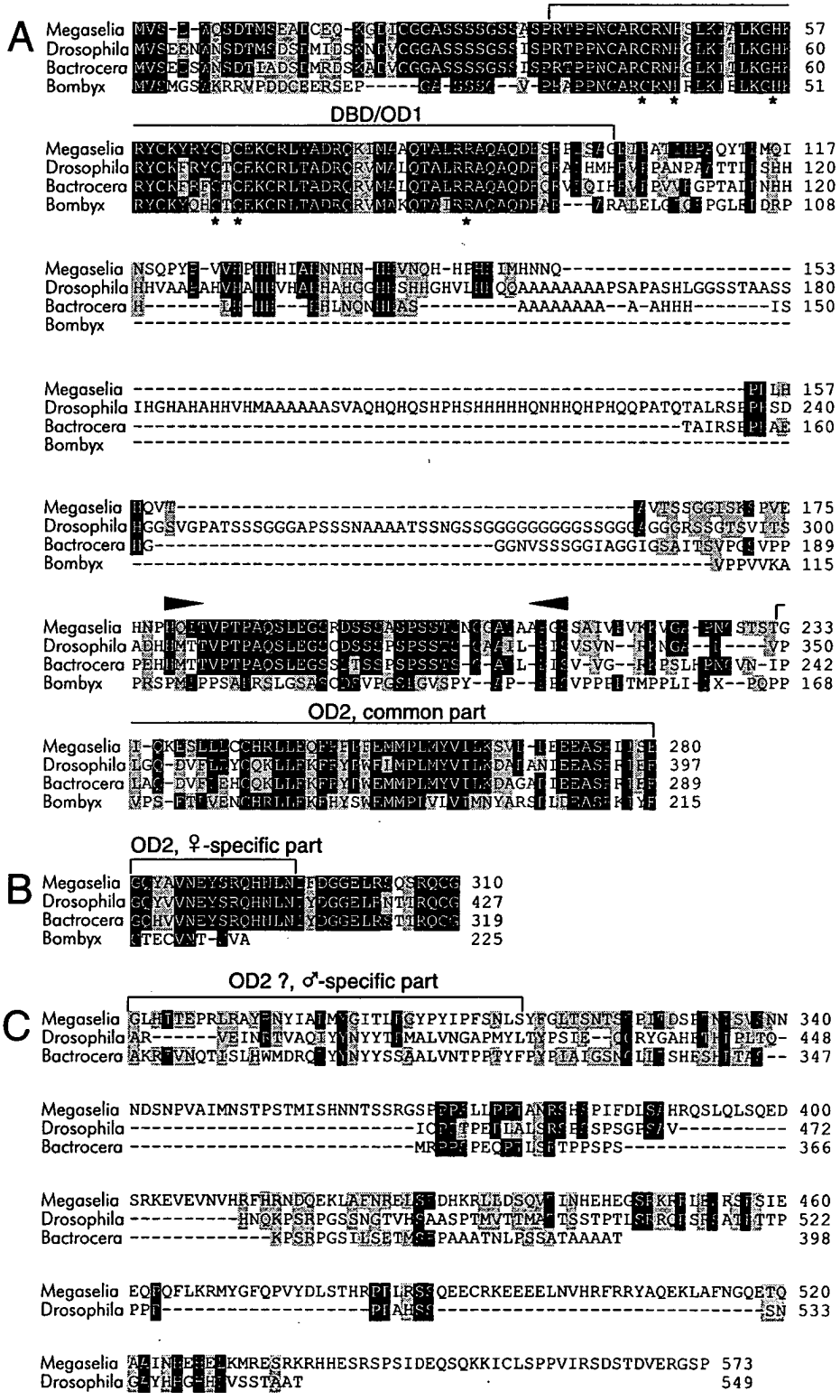
The putative DSX proteins

The female-specific transcript contains an ORF that codes for a putative female-specific protein, DSX^f, of 310 amino acids. It starts in the 5′ common segment (ATG at position 58 in pMSWc178) and ends in the female-specific segment (TGA at position 988 in pMSWc178). The male-specific transcript includes a longer ORF that codes for a putative male-specific protein, DSX^m, of 573 amino acids. The ORF starts at the same site in the 5′ common segment (ATG at position 109 in pMSWc233) as in the female-specific transcript but ends in the 3′ common segment (TAG at position 1828 of pMSWc233).

Megaselia scalaris DSX^f and DSX^m proteins share an N-terminal part of 280 amino acids (common part, Fig. 1B) but differ in a sex-specific C-terminal part, which is short in DSX^f (30 amino acids) and long in DSX^m (293 amino acids). The same pattern of sex-specific protein composition is found in DSX^f and DSX^m from *D. melanogaster* and *B. tryoni* (Burtis and Baker 1989; Shearman and Frommer 1998).

A BLASTX search with pMSWc178 in the nr sequence database of NCBI returned the DSX entries from *B. tryoni* and *D. melanogaster* with highest scores. In a TBLASTN search in the dbEST database, two entries from cDNA libraries of *Bombyx mori*, submitted by K. Mita, M. Morimyo, T. Shimada, K. Okano, and S. Maeda (Chiba, Japan), showed highest similarity. Both were derived from female tissues. BLAST searches with the male-derived pMSWc233 sequence did not return further significant entries.

We selected the DSX^f proteins of the insects *D. melanogaster*, *B. tryoni*, and the translated sequence of one of the *B. mori* cDNAs (Acc. No. AV398350) for a comparison with DSX^f of *M. scalaris*. *M. scalaris* DSX^f has highest similarity with *D. melanogaster* DSX^f (58% identity,



79% similarity; calculations are based on the shorter *M. scalaris* sequence). The protein alignment in Fig. 4A and 4B reveals high similarity among all four proteins at both ends, whereas the central region is barely conserved.

The N-terminal conserved region ranges from the N-terminus up to and including the zinc-finger-like DNA-binding and oligomerization domain (DBD/OD1) of *D. melanogaster* DSX (Erdman and Burtis 1993; An et al.

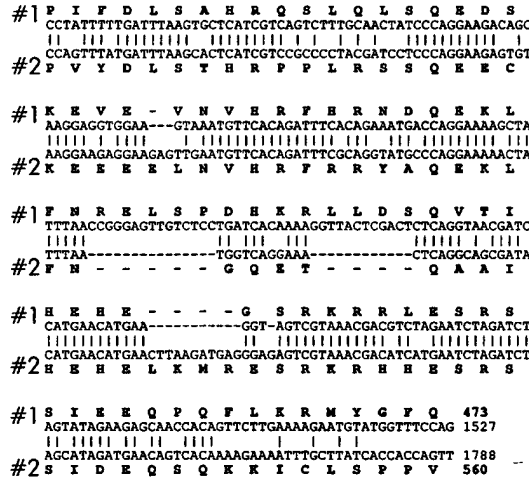


Fig. 5. Alignment of repeats #1 and #2 from the 3′ common segment of *Megaselia scalaris dsx*. Bold letters represent corresponding amino acid sequence from DSX^m.

1996). In DBD/OD1, six amino acids whose replacement by another residue has been shown to abolish DNA-binding activity of the DSX protein (Erdman and Burtis 1993) are conserved in the DSX homologues from *M. scalaris*, *D. melanogaster*, *B. tryoni*, and *B. mori* (Figs. 4A and 7, asterisks).

There are two highly conserved segments in the C-terminal region. The first one, ranging from *M. scalaris* DSX amino acid positions 179 through 216 (Fig. 4A, arrowheads) is functionally undefined yet. It contains a proline- and serine-rich region, but conservation is not restricted to these amino acids. The second segment corresponds to the oligomerization domain 2 (OD2) of *D. melanogaster* DSX (An et al. 1996). It consists of a part that is common to DSX^f and DSX^m (Fig. 4A, OD2, common part) and a female-specific part (Fig. 4B, OD2, ♀-specific part).

It is interesting to note that even the stop signal is conserved; the coding sequence of *dsx*^{f} in all three fly species is terminated by TGATAA, the two stop codons opal and ochre in succession. There is no information on the stop signal in *B. mori*, as the dbEST sequence is truncated at the 3′ end.

In contrast to the female-specific part of DSX^f, the male-specific part of DSX^m shows rather low similarity with the corresponding parts of DSX in *D. melanogaster* and *B. tryoni* (Fig. 4C). They contribute to the male-specific OD2 in *D. melanogaster* and *B. tryoni* but it is unclear how much if any part at all of the male-specific part of DSX^m contributes to the OD2 domain in *M. scalaris* (Fig. 4C, OD2 ?, ♂-specific part).

A common property of the male-specific parts of the three DSX^m homologues is the greater length compared with that of the female-specific parts of DSX^f; the latter ones consist of 30 amino acids while the male-specific part consists of 293 amino acids in *M. scalaris*, 152 amino acids in *D. melanogaster*, and 109 amino acids in *B. tryoni*. The conspicuous length in *M. scalaris* DSX^m is accounted for by two copies of a direct repeat (Fig. 1B; repeat 1 and repeat 2, Fig. 5). The first copy spans 91 amino acids from position 383–473 of *M. scalaris* DSX^m, the second one 87 amino acids, from position 474–560. The two copies are 50% identical and 65% similar at the amino acid level. The similarity is also apparent at the nucleotide level (63% identity).

Intron 3

The inclusion or exclusion of the female-specific exon in *D. melanogaster dsx* depends on the recognition of a weak 3′ splice site with a purine-rich polypyrimidine tract in the preceding intron (intron 3; for convenience, we use the same term for that intron in *M. scalaris*). To retrieve the intron 3 sequence of *M. scalaris*, we performed a nested PCR on genomic DNA with primers DSXK788F/DSXf562R in the first and with DSXK788F/DSXFbr in the second round. A 628-bp fragment of the genomic *M. scalaris dsx* sequence was amplified and cloned in pMSW2473 (Acc. No. AF283697). The fragment contains a short intron of 52 bp (position 100–151 in pMSW2473, Fig. 6).

The general structure of intron 3 in *M. scalaris* is similar to that in *D. melanogaster*, *D. virilis* (Burtis and Baker 1989), and *B. tryoni* (Shearman and Frommer 1998). The 5′ splice sequence is compatible with the 5′ splice sequences of *D. melanogaster*, compiled by Mount et al. (1992). The 3′ splice site, however, appears to be a suboptimal splice acceptor, as only 6 of 12 positions in the polypyrimidine stretch are pyrimidines in *M. scalaris* (Fig. 6). Similarly, in *D. melanogaster* 6 of 12, in *D. virilis* 7 of 12, and in *B. tryoni* 5 of 12 positions are pyrimidines (Burtis and Baker 1989; Shearman and Frommer 1998).

Regulatory elements in the female-specific exon

In *D. melanogaster dsx*, female-specific splicing at the weak 3′ splice site of intron 3 is activated by a *cis*-acting splicing enhancer (dsxRE) and sex-specific *trans*-acting factors (for review, Lopez 1998). The dsxRE is located within the untranslated part of the female-specific exon and consists of "13-nt repeat elements" (Burtis and Baker 1989; Inoue et al. 1992) and a "purine-rich element" (Lynch and Maniatis 1995). In an alignment of the respective *dsx* segments from *M. scalaris* and *D. melanogaster*, we found little nucleotide sequence conservation (data not shown). Nevertheless, *M. scalaris dsx* contains six elements displaying a 10–13 nucleotide (nt) identity with the 13-nt repeat elements (Table 1). Their distance of 247–602 bp downstream from the alternative splice site, is similar to that in *D. melanogaster* (295–566 bp, Burtis and Baker 1989), *D. virilis* (332–458 bp, Hertel et al. 1996), and *B. tryoni* (373–525 bp, Shearman and Frommer 1998).

We found a purine-rich sequence in the female-specific segment of *M. scalaris dsx*^{f} 412 bp downstream of the alternative splice site. It consists of 20 nts, 18 of which are purines (position 1310–1329 in pMSWc178). Purine-rich sequences were found at similar positions in *D. melanogaster* (Lynch and Maniatis 1995), *D. virilis* (Hertel et al. 1996), and *B. tryoni* (Shearman and Frommer 1998). We did not find the direct repeat identified in the purine-rich element of *D. melanogaster*, but this repeat is not conserved in *B. tryoni* or in *D. virilis*.

Molecular phylogeny of the DBD/OD1 domain

A TBLASTN search in the nr database with the zinc-finger-like DNA-binding domain (DBD/OD1) of *M. scalaris* DSX

Fig. 6. The *Megaselia scalaris dsx* intron between the 5′ common segment and the female-specific exon, compared with the corresponding intron 3 of *Drosophila melanogaster, D. virilis* (Burtis and Baker 1989), and *Bactrocera tryoni dsx* (Acc. No. AF040077, Shearman and Frommer 1998). Pyrimidines in the polypyrimidine tract (grey box) are presented in bold letters.

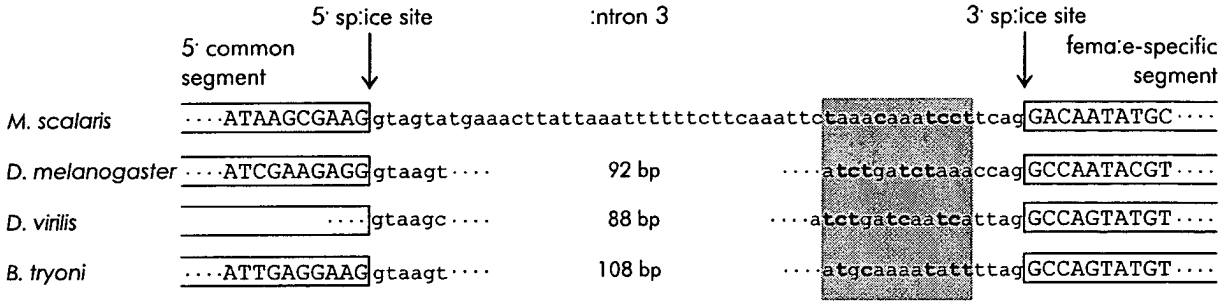
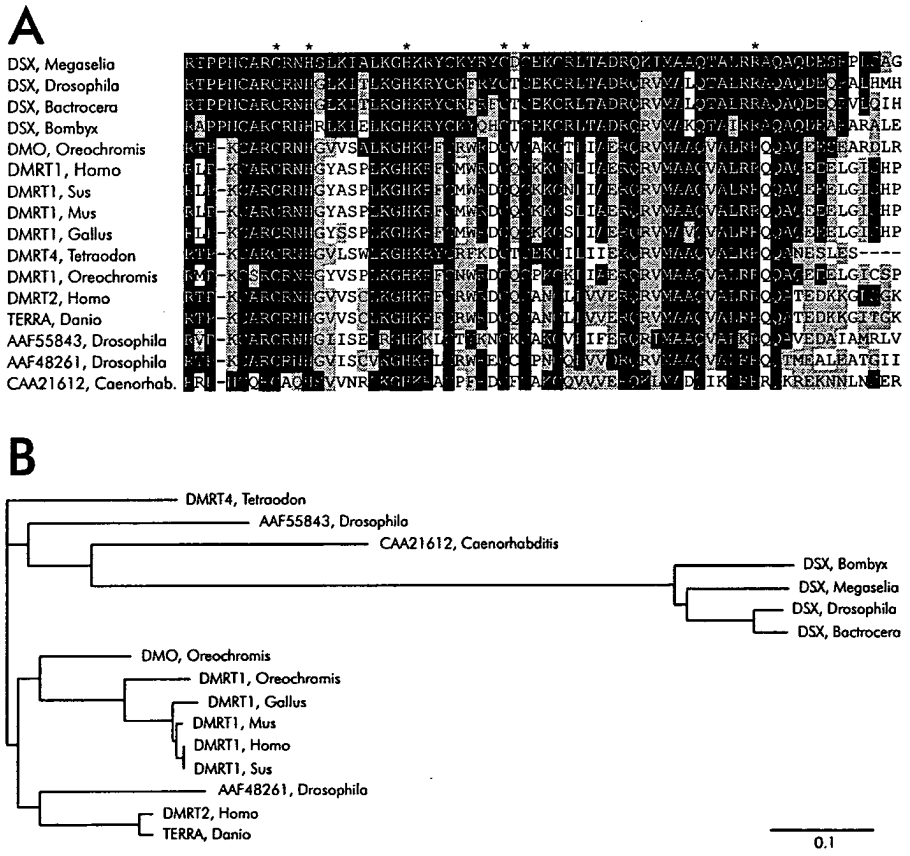
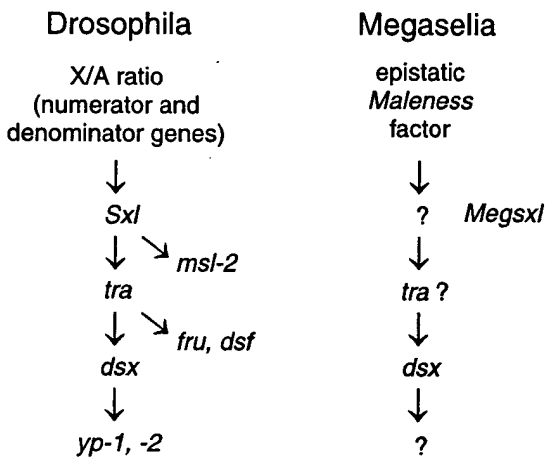


Fig. 7. DNA-binding domains homologous to the DBD/OD1 domain of *Megaselia scalaris* DSX. (A) Alignment of the sequences arranged according to increasing distance from *M. scalaris* DSX. Black boxes, amino acids identical to the *M. scalaris* sequence; shaded boxes, amino acids similar to the *M. scalaris* sequence. Accession numbers: DSX, Drosophila (M25292); DSX, Bactrocera (AF029675); DSX, Bombyx (AV398350); DMO, Oreochromis (AF203490); DMRT1, Oreochromis (AF203489); DMRT1, Gallus (AF123456); DMRT1, Mus (NM_015826); DMRT1, Sus (AF216651); DMRT1, Homo (AJ276801); DMRT4, Tetraodon (AJ251456); DMRT2, Homo (Y19052); TERRA Danio (AF080622); CAA21612, Caenorhabditis (AL032637); AAF48261, Drosophila (AE003492); AAF55843, Drosophila (AE003733). (B) Distance tree. The tree was constructed according to the neighbor-joining method of Saitou and Nei (1987).



returned a series of protein entries with a similar domain. For a phylogenetic comparison with *M. scalaris* DSX, we selected the 14 entries with highest scores, discarding redundancies and sequences with incomplete domains, plus *B. mori* DSX, which had been found in a different database (see above). The amino acid sequence alignment in Fig. 7A shows the DNA-binding domain to be well conserved among these proteins. The dendrogram in Fig. 7B displays the distances between the DBD/OD1 sequences. *D. melanogaster* and vertebrates contribute more than one homologous gene

Fig. 8. Sex-determining pathways in *Drosophila melanogaster* and *Megaselia scalaris*.



to the selected group. They form clusters of paralogous sequences. The four DSX sequences are closely related and form a distinct group. The branch points in this group reflect the phylogenetic relationship among the species as derived from their taxonomic assignment. *Drosophila* and *Bactrocera* belong to the Schizophora, and *Megaselia* to the Aschiza among flies. The silk moth *Bombyx mori* is a representative of a different insect order, Lepidoptera. All other proteins have less similarity to this group and they lack the OD2 domain of DSX (not shown).

Discussion

Orthologues of *dsx*

Transcripts of *M. scalaris dsx* come in two variants: a female-specific form, dsx^f, and a male-specific form, dsx^m, like those from *D. melanogaster* and *B. tryoni* (Burtis and Baker 1989; Shearman and Frommer 1998). The composition of the *dsx* variants, however, differs to some degree. In *D. melanogaster* and *B. tryoni*, dsx^f, and dsx^m each consist of a 5′ common segment and a 3′ sex-specific segment. In *M. scalaris*, dsx^f consists of a common 5′, a female-specific, and a common 3′ segment, whereas dsx^m consists of only the 5′ and the 3′ common segments. This results in longer dsx^f than dsx^m transcripts.

Despite the different compositions of the transcripts, the putative proteins DSX^f and DSX^m of *M. scalaris* are composed similarly to those of *D. melanogaster* and *B. tryoni*. They consist of a common N-terminal part and a sex-specific C-terminal part. This is achieved in dsx^f of *M. scalaris* by a translation stop in the female-specific segment while in dsx^m translation stops in the 3′ common segment.

Sequence conservation is high in the two functional domains DBD/OD1 and OD2 but is not restricted to these regions. DBD/OD1 is a well-conserved domain in a wide range of metazoan species and there are several paralogues apparent in vertebrates and *D. melanogaster*. Molecular phylogeny based on this domain shows *M. scalaris dsx* to be most closely related to *D. melanogaster*, *B. tryoni*, and *B. mori dsx*, indicating that these are orthologous genes.

Sex-specific splicing

In intron 3, we find a poor splice acceptor with a purine-rich polypyrimidine stretch, similar to that present in *D. melanogaster*. Some 250–600 bp downstream in the female-specific segment a purine-rich element and 13-nt repeat elements are conserved. Comparable elements in *D. melanogaster* form *cis*-acting splicing enhancer dsxRE, to which the a *trans*-acting multiprotein complex binds. It consists of TRA2, RBP1, and the female-specific protein TRA, and activates splicing at the weak female-specific 3′ splice site (for review, Lopez 1998). The distance of the enhancer elements from the alternative splice site is critical for their function in *D. melanogaster*. When dsxRE was experimentally moved to less than 150 bp from intron 3, the splicing enhancer function was constitutive and independent of the presence of the TRA2/RBP1/TRA complex (Tian and Maniatis 1994). The distance in *M. scalaris* is appropriate for a corresponding control mechanism based on a *tra*-like step in the sex-determining cascade.

Functional conservation of *dsx*

Drosophila DSX is responsible for transcriptional activation (DSX^f) or suppression (DSX^m) of yolk protein genes, *yp-1* and *yp-2*, in the fat body (for review, Bownes 1994 and references therein; An and Wensink 1995*a*, 1995*b*) but is suspected to regulate sex-specific transcription of more genes (Schütt and Nöthiger 2000). *Drosophila* DSX is known to bind to DNA in the form of a homodimer (An et al. 1996; Erdman et al. 1996). The strong conservation of the DNA-binding and oligomerization domains DBD/OD1 and OD2 indicates conservation of these essential functions in *M. scalaris* DSX^f. A conserved region in front of OD2 is rich in proline and serine residues, but no functional significance of this region is yet known. Some authors suggest that this region mediates transcriptional regulation and (or) protein–protein interaction (Raymond et al. 1999; Yi and Zarkower 1999).

Megaselia scalaris DSX^m contains the same conserved regions as DSX^f with one exception: the sex-specific part of OD2. Due to the overall low degree of conservation in the male-specific part of DSX^m, we can draw no conclusion regarding the extension of OD2 in that protein. In *D. melanogaster*, DSX^f and DSX^m bind to the same site within the promoter region of the yolk protein genes. While DSX^f is activating, DSX^m is repressing transcription. The conspicuous size of the male-specific part of DSX^m in *M. scalaris* may help to fulfill that function by sterically obstructing the binding of activators for the yolk protein gene transcription to the promoter, as suggested for *D. melanogaster* DSX^m by An and Wensink (1995*a*) and Cho and Wensink (1998).

Conservation of the sex-determining mechanism

Results presented in this paper confirm and extend an earlier report of our group on *M. scalaris dsx* (Sievert et al. 1997). From these, a picture of a part of the sex-determining cascade in *M. scalaris* emerges (Fig. 8). Presence or absence of the epistatic (and transposable) *Maleness* factor is the primary sex-determining signal (Traut and Willhoeft 1990; Traut 1994). It exerts its control on an unknown gene in the cascade. *Sxl*, which mediates the sex-determining signal in

D. melanogaster (for review, Schütt and Nöthiger 2000), is not part of *M. scalaris* sex-determining cascade (*Megsxl*, Fig. 8, Sievert et al. 1997; Sievert et al. 2000). The next step in the sex-determining cascade of *Drosophila*, *transformer* (*tra*), has not yet been isolated in *M. scalaris*, but the presence of binding sites for the splice-activating TRA2/RBP1/TRA complex hints at its presence. The next step in the cascade, *dsx*, is conserved in *M. scalaris*. All but one of the structural details considered essential for its proper function as a transmitter of the sex-determining signal are conserved. The one exception is the male-specific component of OD2.

It is obvious that, while primary and secondary sex-determining steps are not conserved, subsequent steps are conserved among flies even when they belong to such distantly related groups as Schizophora (*Drosophila*, *Bactrocera*) and Aschiza (*Megaselia*). The conservation of functional domains in *dsx* of the silk moth, *Bombyx mori*, indicates that this step in the sex-determining pathway is conserved in an even wider range of different insect orders.

It is not clear yet how much of this pathway is conserved in animal groups other than insects. There are intriguing observations from nematodes and vertebrates. *Drosophila* dsx^m rescues *mab-3* mutants in the nematode *C. elegans* (Raymond et al. 1998). The vertebrate gene *Dmrt1/DMRT1* is expressed in the genital ridge of embryos and in testes of adults and probably plays a role in sexual development of vertebrates (Raymond et al. 1998; Raymond et al. 1999; Smith et al. 1999; De Grandi et al. 2000; Guan et al. 2000). Both, DMRT1 and MAB-3, are proteins containing the same type of zinc-finger-like DNA-binding domain as DSX. However, DMRT1 as well as MAB-3 lack the OD2 domain that is characteristic for DSX, and there are other genes in *C. elegans* with higher similarity to DSX (see Fig. 7B). It is obvious that this type of DNA-binding protein plays a wide role in the regulation of sexual development. However, it is not clear whether they play a key part in sex determination, as in the role of *dsx* in insects.

Acknowledgements

We thank Corinna Heide for assistance with the PhosphoImager and Bärbel Kunze for helpful comments on the manuscript, Martin Lipphardt for fruitful discussions, and Jan Schorch for technical assistance.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25**: 3389–3402.

An, W., and Wensink, P.C. 1995*a*. Integrating sex- and tissue-specific regulation within a single *Drosophila* enhancer. Genes Dev. **9**: 256–266.

An, W., and Wensink, P.C. 1995*b*. Three protein binding sites form an enhancer that regulates sex- and fat body-specific transcription of *Drosophila yolk protein* genes. EMBO J. **14**: 1221–1230.

An, W., Cho, S., Ishii, H., and Wensink, P.C. 1996. Sex-specific and non-sex-specific oligomerization domains in both of the *doublesex* transcription factors from *Drosophila melanogaster*. Mol. Cell. Biol. **16**: 3106–3111.

Bedo, D.G., and Foster, G.G. 1985. Cytogenetic mapping of the male-determining region of *Lucilia cuprina* (Diptera: Calliphoridae). Chromosoma, **92**: 344–350.

Blin, N., and Stafford, D.W. 1976. A general method for isolation of high molecular weight DNA from eukaryotes. Nucleic Acids Res. **3**: 2303–2308.

Bownes, M. 1994. The regulation of the *yolk protein* genes, a family of sex differentiation genes in *Drosophila melanogaster*. BioEssays, **16**: 745–752.

Bridges, C.B. 1925. Sex in relation to chromosomes and genes. Am. Nat. **59**: 127–137.

Burtis, K.C., and Baker, B.S. 1989. *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. Cell, **56**: 997–1010.

Cho, S., and Wensink, P.C. 1998. Linkage between oligomerization and DNA binding in *Drosophila doublesex* proteins. Biochemistry, **37**: 11 301 – 11 308.

Cline, T.W., and Meyer, B.J. 1996. Vive la difference: Males vs. females in flies vs. worms. Annu. Rev. Genet. **30**: 637–702.

De Grandi, A., Calvari, V., Bertini, V., Bulfone, A., Peverali, G., Camerino, G., Borsani, G., and Guioli, S. 2000. The expression pattern of a mouse *doublesex*-related gene is consistent with a role in gonadal differentiation. Mech. Dev. **90**: 323–326.

Dübendorfer, A., Hilfiker-Kleiner, D., and Nöthiger, R. 1992. Sex determination mechanisms in dipteran insects: the case of *Musca domestica*. Semin. Dev. Biol. **3**: 349–356.

Erdman, S.E., and Burtis, K.C. 1993. The *Drosophila doublesex* proteins share a novel zinc finger related DNA binding domain. EMBO J. **12**: 527–535.

Erdman, S.E., Chen, H.J., and Burtis, K.C. 1996. Functional and genetic characterization of the oligomerization and DNA binding properties of the *Drosophila doublesex* proteins. Genetics, **144**: 1639–1652.

Guan, G., Kobayashi, T., and Nagahama, Y. 2000. Sexually dimorphic expression of two types of DM (*doublesex/mab-3*)-domain genes in a teleost fish, the tilapia (*Oreochromis niloticus*). Biochem. Biophys. Res. Commun. **272**: 662–666.

Henikoff, S., and Henikoff, J.G. 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. U.S.A. **89**: 10 915 – 10 919.

Hertel, K.J., Lynch, K.W., Hsiao, E.C., Liu, E.H., and Maniatis, T. 1996. Structural and functional conservation of the *Drosophila doublesex* splicing enhancer repeat elements. RNA (New York), **2**: 969–981.

Inoue, K., Hoshijima, K., Higuchi, I., Sakamoto, H., and Shimura, Y. 1992. Binding of the *Drosophila transformer* and *transformer-2* proteins to the regulatory elements of *doublesex* primary transcript for sex-specific RNA processing. Proc. Natl. Acad. Sci. U.S.A. **89**: 8092–8096.

Kuwabara, P.E. 1999. Developmental genetics of *Caenorhabditis elegans* sex determination. Curr. Top. Dev. Biol. **41**: 99–132.

Lopez, A.J. 1998. Alternative splicing of pre-mRNA: Developmental consequences and mechanisms of regulation. Annu. Rev. Genet. **32**: 279–305.

Lynch, K.W., and Maniatis, T. 1995. Synergistic interactions between two distinct elements of a regulated splicing enhancer. Genes Dev. **9**: 284–293.

Mainx, F. 1964. The genetics of *Megaselia scalaris* Leow (Phroidae): A new type of sex determination in Diptera. Am. Nat. **XCVIII**: 415–430.

Mainx, F. 1966. Die Geschlechtsbestimmung bei *Megaselia scalaris* Loew (Phoridae). Z. Vererbungsl. **98**: 49–60.

Meise, M., Hilfiker-Kleiner, D., Dübendorfer, A., Brunner, C., Nöthiger, R., and Bopp, D. 1998. *Sex-lethal*, the master sex-

determining gene in *Drosophila*, is not sex-specifically regulated in *Musca domestica*. Development, **125**: 1487–1494.

Mount, S.M., Burks, C., Hertz, G., Stormo, G.D., White, O., and Fields, C. 1992. Splicing signals in *Drosophila*: Intron size, information content, and consensus sequences. Nucleic Acids Res. **20**: 4255–4262.

Müller-Holtkamp, F. 1995. The *Sex-lethal* gene homologue in *Chrysomya rufifacies* is highly conserved in sequence and exon-intron organization. J. Mol. Evol. **41**: 467–477.

Page, R.D. 1996. TREEVIEW: An application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**: 357–358.

Raymond, C.S., Shamu, C.E., Shen, M.M., Seifert, K.J., Hirsch, B., Hodgkin, J., and Zarkower, D. 1998. Evidence for evolutionary conservation of sex-determining genes. Nature, **391**: 691–695.

Raymond, C.S., Kettlewell, J.R., Hirsch, B., Bardwell, V.J., and Zarkower, D. 1999. Expression of *Dmrt1* in the genital ridge of mouse and chicken embryos suggests a role in vertebrate sexual development. Dev. Biol. **215**: 208–220.

Saccone, G., Peluso, I., Artiaco, D., Giordano, E., Bopp, D., and Polito, L.C. 1998. The *Ceratitis capitata* homologue of the *Drosophila* sex-determining gene *Sex-lethal* is structurally conserved, but not sex-specifically regulated. Development, **125**: 1495–1500.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**: 406–425.

Sambrook, J., Fritsch, E.F., and Maniatis, T. 1989. Molecular Cloning: A Laboratory Manual. 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Schütt, C., and Nöthiger, R. 2000. Structure, function and evolution of sex-determining systems in Dipteran insects. Development, **127**: 667–677.

Shearman, D.C., and Frommer, M. 1998. The *Bactrocera tryoni* homologue of the *Drosophila melanogaster* sex-determination gene *doublesex*. Insect Mol. Biol. **7**: 355–366.

Sievert, V., Kuhn, S., and Traut, W. 1997. Expression of the sex determining cascade genes *Sex-lethal* and *doublesex* in the phorid fly *Megaselia scalaris*. Genome, **40**: 211–214.

Sievert, V., Kuhn, S., Paululat, A., and Traut, W. 2000. Sequence conservation and expression of the *Sex-lethal* homologue in the fly *Megaselia scalaris*. Genome, **43**: 382–390.

Smith, C.A., McClive, P.J., Western, P.S., Reed, K.J., and Sinclair, A.H. 1999. Conservation of a sex-determining gene. Nature, **402**: 601–602.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**: 4673–4680.

Tian, M., and Maniatis, T. 1994. A splicing enhancer exhibits both constitutive and regulated activities. Genes Dev. **8**: 1703–1712.

Traut, W. 1994. Sex determination in the fly *Megaselia scalaris*, a model system for primary steps of sex chromosome evolution. Genetics, **136**: 1097–1104.

Traut, W., and Willhoeft, U. 1990. A jumping sex determining factor in the fly *Megaselia scalaris*. Chromosoma, **99**: 407–412.

Ullerich, F.H. 1984. Analysis of sex determination in the monogenic blowfly *Chrysomya rufifacies* by pole cell transplantation. Mol. Gen. Genet. **193**: 479–487.

Willhoeft, U., and Franz, G. 1996. Identification of the sex-determining region of the *Ceratitis capitata* Y chromosome by deletion mapping. Genetics, **144**: 737–745.

Willhoeft, U., and Traut, W. 1990. Molecular differentiation of the homomorphic sex chromosomes in *Megaselia scalaris* (Diptera) detected by random DNA probes. Chromosoma, **99**: 237–242.

Yi, W., and Zarkower, D. 1999. Similarity of DNA binding and transcriptional regulation by *Caenorhabditis elegans* MAB-3 and *Drosophila melanogaster* DSX suggests conservation of sex determining mechanisms. Development, **126**: 873–881.

RNA (1996), 2:969-981. Cambridge University Press. Printed in the USA.

Structural and functional conservation of the *Drosophila doublesex* splicing enhancer repeat elements

KLEMENS J. HERTEL, KRISTEN W. LYNCH, EDWARD C. HSIAO, ERIC H.-T. LIU, and TOM MANIATIS
Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138, USA

ABSTRACT

We have compared the RNA sequences and secondary structures of the *Drosophila melanogaster* and *Drosophila virilis doublesex* (*dsx*) splicing enhancers. The sequences of the two splicing enhancers are highly divergent except for the presence of nearly identical 13-nt repeat elements (six in *D. melanogaster* and four in *D. virilis*) and a stretch of nucleotides at the 5′ and 3′ ends of the enhancers. In vitro RNA structure probing of the two enhancers revealed that the 13-nt repeats are predominantly single-stranded. Thus, both the primary sequences and single-stranded nature of the repeats are conserved between the two species. The significance of the primary sequence conservation was demonstrated by showing that the two enhancers are functionally interchangeable in Tra-/Tra2-dependent in vitro splicing. In addition, inhibition of splicing enhancer activity by antisense oligonucleotides complementary to the repeats demonstrated the importance of the conserved single-stranded structure of the repeats. In vitro binding studies revealed that Tra2 interacts with each of the *D. melanogaster* repeat elements, except for repeat 2, with affinities that are indistinguishable, whereas Tra binds nonspecifically to the enhancer. Taken together, these observations indicate that the organization of sequences within the *dsx* splicing enhancers of *D. melanogaster* and *D. virilis* results in a structure in which each of the repeat elements is single-stranded and therefore accessible for specific recognition by the RNA-binding domain of Tra2.

Keywords: phylogeny; regulated splicing; RNA/protein interactions; RNA structure

INTRODUCTION

Sex-specific alternative splicing of the *Drosophila melanogaster doublesex* (*dsx*) pre-mRNA requires the regulatory proteins Transformer (Tra) and Transformer 2 (Tra2), and a Tra- and Tra2-dependent splicing enhancer (the *doublesex* repeat element *dsx*RE) that is located 300-nt downstream of the female-specific 3′ splice site (for review see Baker, 1989; Maniatis, 1991). Tra is produced exclusively in females by the sex-specific splicing of Tra pre-mRNA, whereas Tra2 is expressed in both males and females (Boggs et al., 1987; Amrein et al., 1988). As shown in Figure 1, *dsx* pre-mRNA contains six exons: three common exons (exons 1–3), a female-specific exon (exon 4), and two male-specific exons (exons 5 and 6). In males, exon 3 is joined to exon 5 to produce an mRNA containing exons 1, 2, 3, 5, and 6. In females, exon 3 is joined to exon 4 to produce an mRNA containing exons 1, 2, 3, and 4 (Burtis & Baker, 1989). The female-specific 3′ splice site in intron 3 deviates significantly from the consensus 3′ splice site, so it is not recognized by the splicing machinery in the absence of Tra and Tra2, or in the absence of the *dsx*RE (Burtis & Baker, 1989; Hedley & Maniatis, 1991; Hoshijima et al., 1991; Ryner & Baker, 1991; Tian & Maniatis, 1992, 1993; Zuo & Maniatis, 1996).

The *dsx*RE is a 270-nt regulatory element that contains six 13-nt repeat sequences, and a purine-rich element (PRE) located between repeats 5 and 6 (Burtis & Baker, 1989; Lynch & Maniatis, 1995). The presence of both types of elements is required for efficient Tra/Tra2-dependent use of the female-specific 3′ splice site. Tra and Tra2 bind to the *dsx*RE and facilitate the recruitment of splicing factors to the weak female-specific 3′ splice site (Hedley & Maniatis, 1991; Inoue et al., 1992; Tian & Maniatis, 1992, 1993; Zuo & Maniatis, 1996). Thus, in females, the splicing machinery is preferen-

Reprint requests to: Tom Maniatis, Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA.

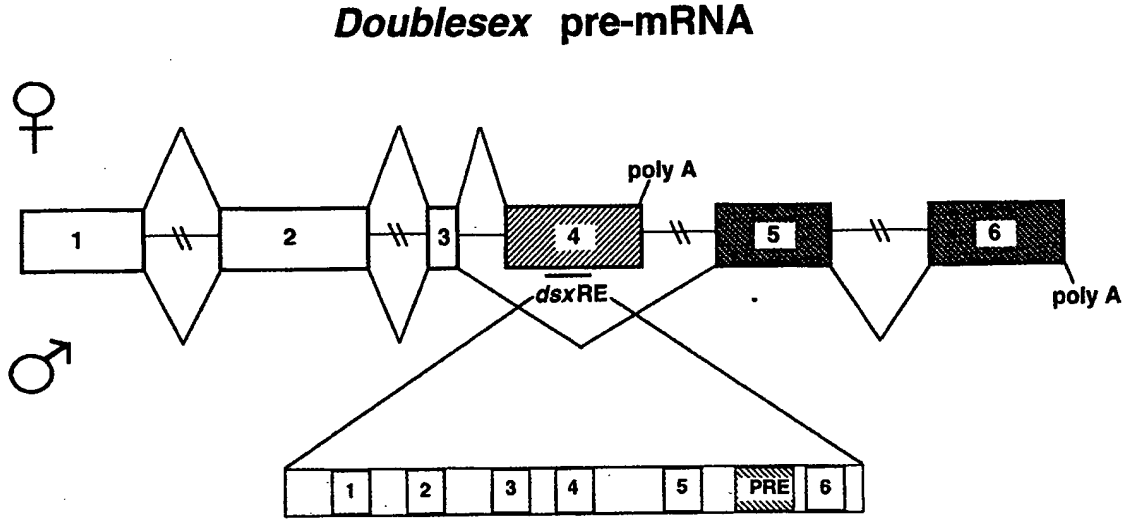


FIGURE 1. Sex-specific alternative splicing pattern of *dsx* pre-mRNA and the *dsx* splicing enhancer (*dsx*RE). Top: Open boxes represent the common exons 1, 2, and 3, the light hatched box represents the female-specific exon 4, and the dark hatched boxes represent the male-specific exons 5 and 6. The sex-specific splicing pattern is illustrated by the lines above (female) and below (male) the pre-mRNA. Sites of cleavage and polyadenylation are labeled poly A. Bottom: Enlargement shows the organization of the *D. melanogaster dsx*RE comprised of six 13-nt repeat elements (boxes 1–6) and a PRE located between repeats 5 and 6.

tially directed to an intrinsically weak splice site recognition signal.

In vitro splicing experiments have shown that, in addition to Tra and Tra2, one or more members of the SR (serine/arginine) family of general splicing factors are required for *dsx*RE-dependent recognition of the female-specific 3′ splice site (Tian & Maniatis, 1993). UV-crosslinking experiments and the characterization of affinity-purified enhancer complexes have shown that Tra and Tra2 recruit SR proteins to the *dsx*RE to form a multicomponent splicing enhancer complex (Tian & Maniatis, 1992, 1993). Both Tra and Tra2 contain SR domains, and are therefore considered members of the superfamily of SR-containing splicing factors (for review see Fu, 1995). Although Tra2 and SR proteins contain an RNA recognition motif (RRM) (Bandziulis et al., 1989), Tra is lacking a known RNA-binding domain. Binding studies in HeLa cell nuclear extracts or with purified recombinant Tra, Tra2, and SR proteins revealed that Tra2 binds with significant specificity to the *dsx*RE, whereas Tra exhibits low or no specificity (Hedley & Maniatis, 1991; Inoue et al., 1992; Tian & Maniatis, 1992; Lynch & Maniatis, 1995, 1996). In addition, these studies showed that Tra, Tra2, and SR proteins bind cooperatively to the *dsx*RE. Recent UV-crosslinking experiments have shown that Tra and Tra2 recruit a specific SR protein to the *dsx* repeats (Lynch & Maniatis, 1996). This SR protein binds to the 5′ ends of the repeats, whereas Tra2 binds to the middle and 3′ regions. Tra is an essential component of this heterotrimeric complex, but does not appear to directly contact RNA.

These observations, in conjunction with studies showing that these proteins interact with each other through their SR domains (Wu & Maniatis, 1993; Amrein et al., 1994), suggest that a complex involving multiple specific protein–protein and protein–RNA interactions is assembled on the *dsx*RE.

Tra2 and SR proteins contact RNA through their RRMs, but relatively little is known about the structure of the binding sites in the *dsx*RE. Recently, the three-dimensional structure of a complex between the RNA-binding protein U1A and its binding site in U1 snRNA was determined (Oubridge et al., 1994). U1A, which contains an RRM similar to that present in SR proteins and Tra2, specifically interacts with the single-stranded region of a hairpin structure formed by U1 snRNA. The single-stranded region provides a surface for extensive interactions between the protein and exposed nucleotides. These results and other studies suggest that interactions with single-stranded RNA might be a general mode of recognition for the RRM domain/RNA complex (Nagai et al., 1995).

Although the *dsx*RE is the only regulated splicing enhancer thus far characterized, a number of constitutive splicing enhancers have been described (Sun et al., 1993; Watakabe et al., 1993; Dominski & Kole, 1994; von Oers et al., 1994; Ramchatesingh et al., 1995). Most of these elements are short (approximately 10 nt) purine-rich sequences, but a few are pyrimidine-rich. None thus far reported resemble the *dsx* repeat sequences. Both types of enhancers require SR proteins for their activities, but only the *dsx*RE requires Tra and Tra2. The primary difference between constitutive en-

hancers and the *dsx*RE is that the former can function only within 100 nt of the affected 3′ splice site, whereas the latter can function at least 1,000 nt away (Tian & Maniatis, 1994). This ability to function at a distance may be due, in part, to the complex organization of the *dsx*RE, and to Tra and Tra2, which may function to promote enhancer complex assembly and stability. Consistent with both of these possibilities is the observation that individual *dsx* repeats or the PRE function as constitutive (Tra- and Tra2-independent) splicing enhancers in vitro when located within 100 nt of the female-specific 3′ splice site (Lynch & Maniatis, 1995). Although individual repeats can function as Tra- and Tra2-dependent enhancers at a distance (Hoshijima et al., 1991), maximal splicing efficiency in vitro requires the combination of multiple repeats, the PRE and Tra and Tra2 (Tian & Maniatis, 1994; Lynch & Maniatis, 1995).

The unique organization of the *dsx*RE suggests the interesting possibility that this arrangement of sequences results in the formation of a specific secondary and tertiary structure that is required for optimal Tra- and Tra2-dependent splicing. To investigate this possibility, we have compared the sequence of the *D. melanogaster dsx*RE with the corresponding sequence from the distantly related *D. virilis*. In addition, we have conducted chemical and enzymatic RNA probing experiments to investigate the secondary structure of the *dsx*REs from both species. This structural information was then used to optimize computer-assisted RNA-folding predictions. These studies revealed significant phylogenetic conservation of the primary sequence of the repeat elements, and the secondary structure of the complete element.

We also conducted RNase footprinting experiments with purified Tra and Tra2 proteins and both *dsx*REs to investigate specific protein-RNA interactions in the complex. Based on these assays, we found that Tra2 binds to each of the repeat elements with affinities that are indistinguishable, whereas Tra binds nonspecifically to the *dsx*RE. We conclude that the organization of the *dsx*RE results in the formation of an RNA secondary structure in which each of the repeats is present as single-stranded RNA that is recognized specifically by the RRMs of both Tra2 and SR proteins.

RESULTS

Comparison of the *D. melanogaster dsx*RE and the corresponding region in the *D. virilis dsx* pre-mRNA

The female-specific fourth exon of *D. virilis* was identified from a genomic DNA library (Newfeld et al., 1991) using the PCR-amplified third intron of *D. virilis* as the hybridization probe (intron sequence from Burtis & Baker, 1989). The DNA sequence of this exon was determined and compared to the corresponding region of the *D. melanogaster dsx* gene. An alignment of the two sequences revealed several highly conserved regions between the two species (Fig. 2). The first 100 nt of the fourth exon are 90% identical between the two species, encoding only 1 different amino acid of the 30 translated amino acids (98% homology on the amino acid level). Because this region contains the coding sequence for the carboxy terminus of the female-specific *dsx* protein, this high degree of sequence conservation is not unexpected. Previously, an interspecific nucleotide sequence comparison of the *Drosophila* hsp82 gene demonstrated that the coding regions of the distantly related *D. melanogaster* and *D. virilis* species are 90% homologous at the DNA level and 97–99% identical at the amino acid level (Blackman & Meselson, 1986). In contrast, little or no sequence conservation was observed in the intron or the nontranslated exon 1 sequences of hsp82. Consistent with this observation, the conservation of noncoding sequences of exon 4 in *dsx* is weak, except at the 5′ and 3′ ends of the *dsx*RE and the repeat sequences. As shown in Figure 2, there are two regions of approximately 30 nt at the 5′ and 3′ ends of the *dsx*RE that are highly conserved. Although these regions in the *D. melanogaster dsx*RE are not required for maximal levels of Tra- and Tra2-dependent splicing in vitro (K.W. Lynch & T. Maniatis, unpubl.), they may be required for a regulatory function in vivo.

By introducing gaps into both *dsx*RE sequences, it is possible to align the 13-nt repeat sequences of the *D. melanogaster dsx*RE with nearly identical sequences in the *D. virilis* exon 4. In agreement with a recent study (Heinrichs & Baker, 1995), six repeat elements are present in the *D. melanogaster dsx*RE, whereas only four such repeats were found in the enhancer region of *D. virilis*. In addition, the nucleotide composition of the 13-nt repeat elements varies slightly in *D. melanogaster* (Burtis & Baker, 1989), whereas all of the repeats in *D. virilis* are identical to each other and to the predominant repeat sequence in *D. melanogaster*. Thus, the repeat sequences are highly conserved between the two species, but the surrounding sequences are nearly random, suggesting that specific sequences are not required for *dsx*RE function outside of the repeats.

The *D. virilis dsx*RE does not contain a sequence that is identical to the *D. melanogaster* PRE. However, there is a purine-rich sequence located immediately downstream of the fourth repeat in the *D. virilis dsx*RE. The factors that bind to this purine-rich sequence are similar to the factors associated with the *D. melanogaster* PRE, and both purine-rich sequences are sufficient to act as a constitutive enhancer when in close proximity to the weak 3′ splice site (K.W. Lynch & T. Maniatis, unpubl.). Thus, the function, but not the exact sequence, of the PRE may be conserved between the two species.

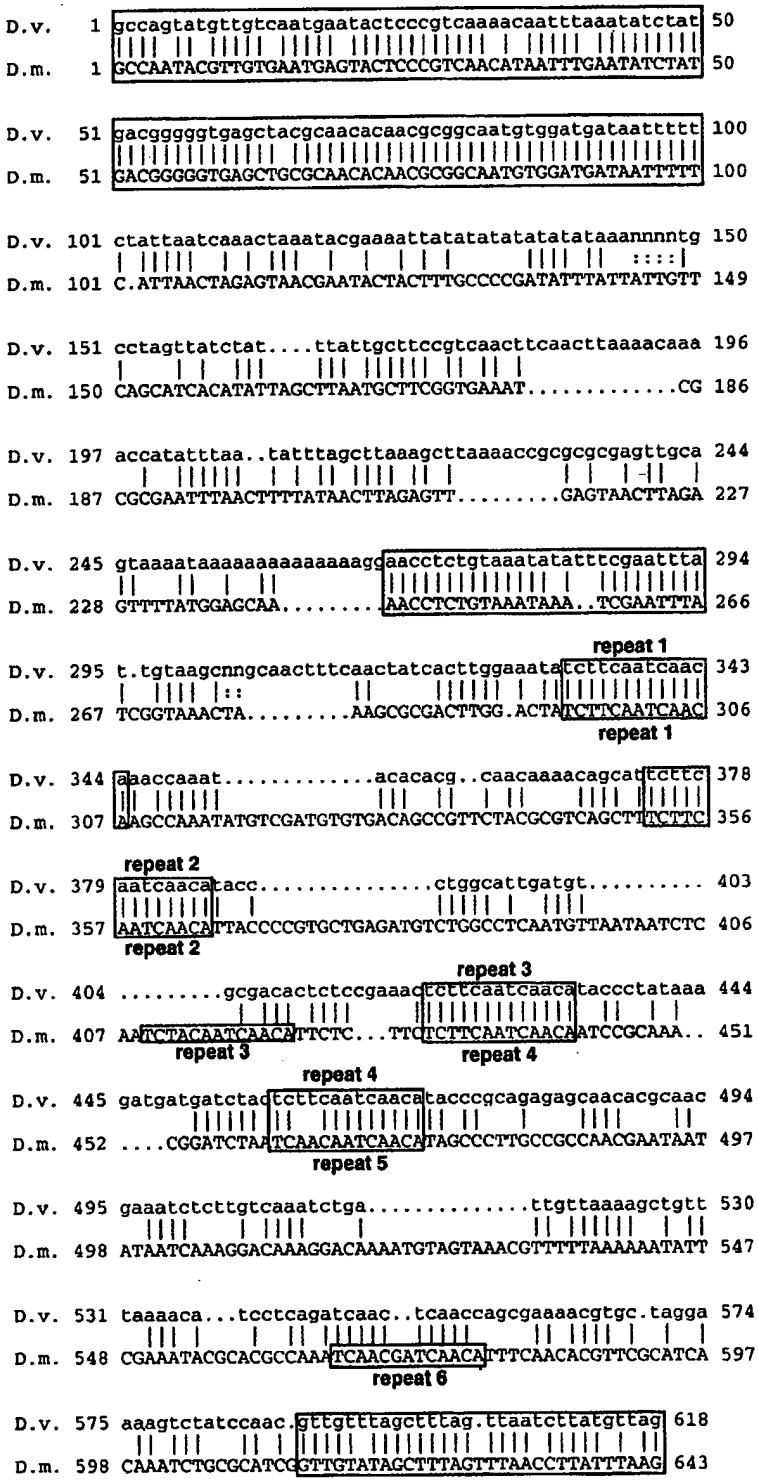


FIGURE 2. Sequence alignment of the female-specific fourth exon from *D. melanogaster* and the distantly related *D. virilis*. Shown are the first 650 nt of the fourth exon, including the *dsx*RE. Vertical bars indicate identical nucleotides. Alignment was accomplished by using the Sequence Analysis Software Package, version 7.2, from the Genetics Computer Group. The 13-nt repeat elements identified in each species are labeled and boxed. Other boxed regions are regions of high sequence homology.

Comparison of the *D. melanogaster* and *D. virilis* splicing enhancer activities in HeLa cell nuclear extracts

To investigate the biological significance of the sequence conservation between the *D. melanogaster* and *D. virilis dsx*REs, we compared their activities in an in vitro splicing assay. To accomplish this, a pre-mRNA substrate (D2) was constructed in which the *dsx*RE of *D. melanogaster* was replaced by the enhancer region of *D. virilis* containing only four repeats. Consistent with earlier studies, splicing of substrate D1 was observed

only in the presence of Tra and Tra2 (Fig. 3). Similarly, the in vitro splicing activity of the *D. virilis dsx*RE required *D. melanogaster* Tra and Tra2, and the concentrations of these proteins required for maximal splicing efficiency are indistinguishable from the concentrations required for substrate D1. Thus, the *dsx*RE from both species can substitute functionally for each other in in vitro splicing assays.

The structural analysis of the *dsx* enhancer described below was conducted with RNAs containing portions of the fourth exon, but lacking the adjacent intron. It

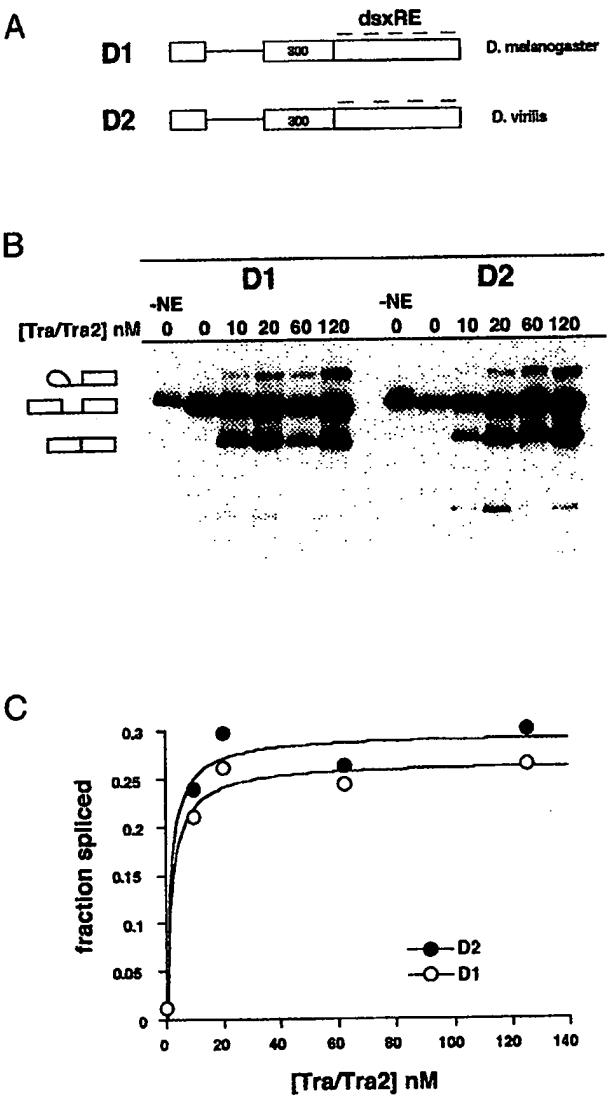


FIGURE 3. *dsx*RE from *D. virilis* can functionally substitute for the *dsx*RE in *D. melanogaster*. **A:** Both oligonucleotides D1 and D2 share part of exon 3 and the entire regulated intron derived from *D. melanogaster*. Substrate D2 contains the *D. virilis dsx*RE in place of the *D. melanogaster dsx*RE. **B:** Splicing efficiencies of D1 and D2 are compared as a function of Tra/Tra2 concentration. **C:** Quantitation of the data in B. At the Tra/Tra2 concentrations used, the splicing efficiency of D1 (open circles) is indistinguishable from the splicing efficiency observed for D2 (closed circles).

is therefore important to demonstrate that the isolated *dsx*RE can form a specific regulatory complex. In fact, several lines of evidence show that *dsx*RE RNA fragments are capable of specifically binding to or competing for Tra/Tra2. For example, an isolated *dsx*RE containing all six repeat elements can specifically inhibit the splicing of D1 (Tian & Maniatis, 1993) and it interacts specifically with Tra and Tra2 in nuclear extracts (Tian & Maniatis, 1992). Similar competition experiments conducted with isolated enhancer elements used here are in agreement with the results of Tian and Maniatis (1993). The titration of enhancer competitor RNA reduced the splicing efficiency of D1 dramatically, whereas the presence of a nonspecific competitor at the same concentration had no significant effect (data not shown). In addition, the splicing efficiency of the Tra/Tra2-independent β-globin pre-mRNA was unaffected by either specific or nonspecific competitor RNA. Thus, the isolated *dsx*REs inhibit *dsx* splicing specifically by competing for *trans*-acting factors that are not components of the basic spliceosome, but are essential for the Tra/Tra2-dependent recruitment of the spliceosome to the *dsx* pre-mRNA. It is therefore reasonable to argue that enhancer elements in the absence of any splice sites are capable of binding to or competing for regulatory factors required for female-specific splicing of *dsx* pre-mRNA.

Secondary structure analysis of the *D. melanogaster* and *D. virilis dsx*REs

Direct enzymatic and chemical in vitro structure probing was used to identify regions within the enhancer elements that are single stranded or involved in secondary or higher-order interactions (Knapp, 1989; Krol & Carbon, 1989). The digestion pattern of the *D. melanogaster dsx*RE RNA revealed that the 13-nt repeat elements are predominantly in a single-stranded configuration (Fig. 4). Our analysis also identified several base paired regions with different sensitivities to RNase V1, an enzyme that specifically recognizes nucleotides involved in base pairing. Similarly, absence of RNase T1 digestion is indicative of guanosine residues that are not accessible for modification. These nucleotides are thought to be in the immediate vicinity of or directly involved in higher-order structural arrangements.

Three different permutations of the *D. melanogaster* enhancer region (R1-6, R2-5PRE, R2-6) were used to address whether the nucleotides 5′ or 3′ from the repeat elements influence the folding pattern. Neither the removal of the first nor the last repeat element of the *dsx*RE changed the digestion pattern (data not shown). Thus, the in vitro folding of the enhancer element does not depend on interactions between the 5′ and 3′ ends of the RNA.

In order to substantiate the secondary structure information of the *D. melanogaster* enhancer region ob-

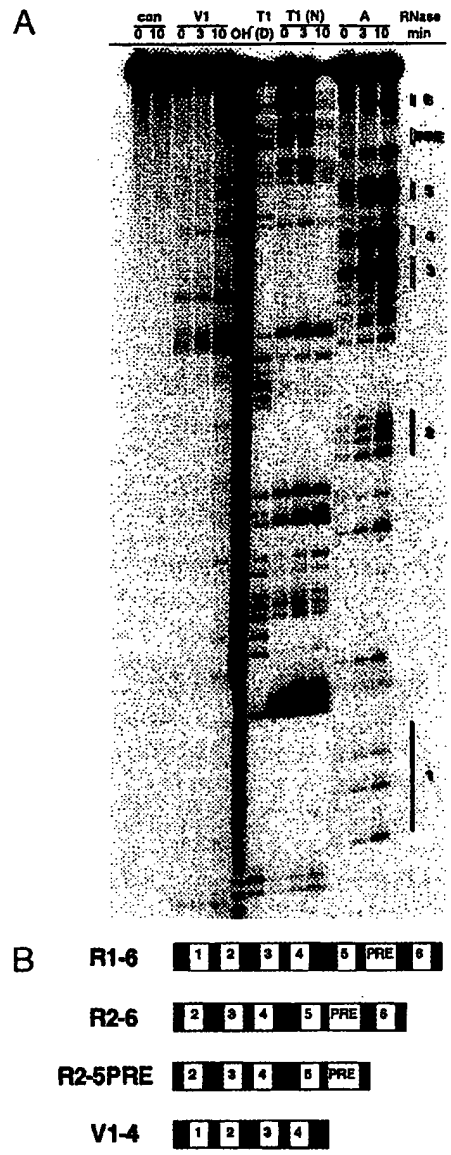


FIGURE 4. In vitro structure probing of the *dsx*RE. **A:** RNase digestion pattern of R1-6. Lanes: con, no RNase added; V1, RNase VI digestion; OH^-, alkaline hydrolysis ladder; T1 (D), RNase T1 digestion at denaturing conditions; T1(N), RNase T1 digestion at assay conditions; A, RNase A digestion. Locations of repeats 1-6 and the PRE are indicated on the right side. **B:** Summary of the various oligonucleotides used for the *dsx*RE structure probing. R1-6 contains all repeats present in the *dsx*RE of *D. melanogaster* and the PRE; R2-5PRE contains repeats 2-5 and the PRE; R2-6 contains repeats 2-6, including the PRE, and V1-4 contains the *dsx*RE from *D. virilis* containing all four repeat elements.

tained by nuclease digestion, the RNA was treated with the single-stranded chemical modifier DMS followed by primer extension (Krol & Carbon, 1989). DMS methylates adenine and cytosine residues at the N-1 and N-3 positions, respectively, with some preference for adenine. Residues that interact with other nucleotides through N-1- or N-3-mediated hydrogen bonding are not accessible for the chemical modification. The results of a series of DMS methylation experiments are summarized in Figure 5A. As observed in the RNase digestion experiments, the 13-nt repeat elements are very accessible to chemical modification and are therefore in a predominantly single-stranded conformation. By contrast, the PRE appears to be in a predominantly base-paired configuration.

Similar experiments were conducted with the *D. virilis dsx*RE. As with the *D. melanogaster dsx*RE, the 13-nt repeat elements of the *D. virilis* RNA are predominantly in single-stranded regions. Only one of the four repeats appears to be involved in some secondary structure (Fig. 5B).

Computer-assisted folding of the enhancer region

The enhancer regions of *D. melanogaster* and *D. virilis* were folded using the MFOLD and PLOTFOLD application programs (version 7.2) from the Genetics Computer Group, University of Wisconsin Biotechnology Center. With the availability of biochemical structure data collected in the RNase and chemical modification experiments described above, the folding of several nucleotides within each RNA was prevented prior to the application of the program. Because MFOLD can generate and analyze suboptimal structures, a representative secondary structure depiction was chosen from a series of optimal and suboptimal structures based on its agreement with the remaining experimental data. Figure 6 illustrates structure representations for the enhancer regions of *D. melanogaster* and *D. virilis* that best fit the biochemical data. With the exception of repeat 2 in each enhancer, all of the 13-nt repeats are single stranded. In the representations, approximately 40% of each RNA is involved in Watson-Crick base pairing. This is a relatively low percentage compared to the well-defined RNA structures within the ribosomal RNAs (Noller, 1984), but similar to the extent observed for U2 snRNA in *Tetrahymena* (Zaug & Cech, 1995).

Tra/Tra2-dependent in vitro footprinting of the enhancer region

Previous studies have shown that Tra2 can bind specifically to the *dsx*RE (Hedley & Maniatis, 1991) or a short oligonucleotide comprised of two repeat elements (Inoue et al., 1992), but the site of this interaction is not known. Therefore, we conducted in vitro footprinting studies to identify Tra2 binding sites within the *dsx*RE. An initial binding specificity screen was established to evaluate the binding specificity of Tra or Tra2 to the enhancer probe. In this screen, a mixture of 5′ end-labeled oligonucleotides cleaved at a single residue by alkaline hydrolysis was mixed with increasing concentrations of Tra or Tra2, and the bound molecules were recovered by retention on nitrocellulose

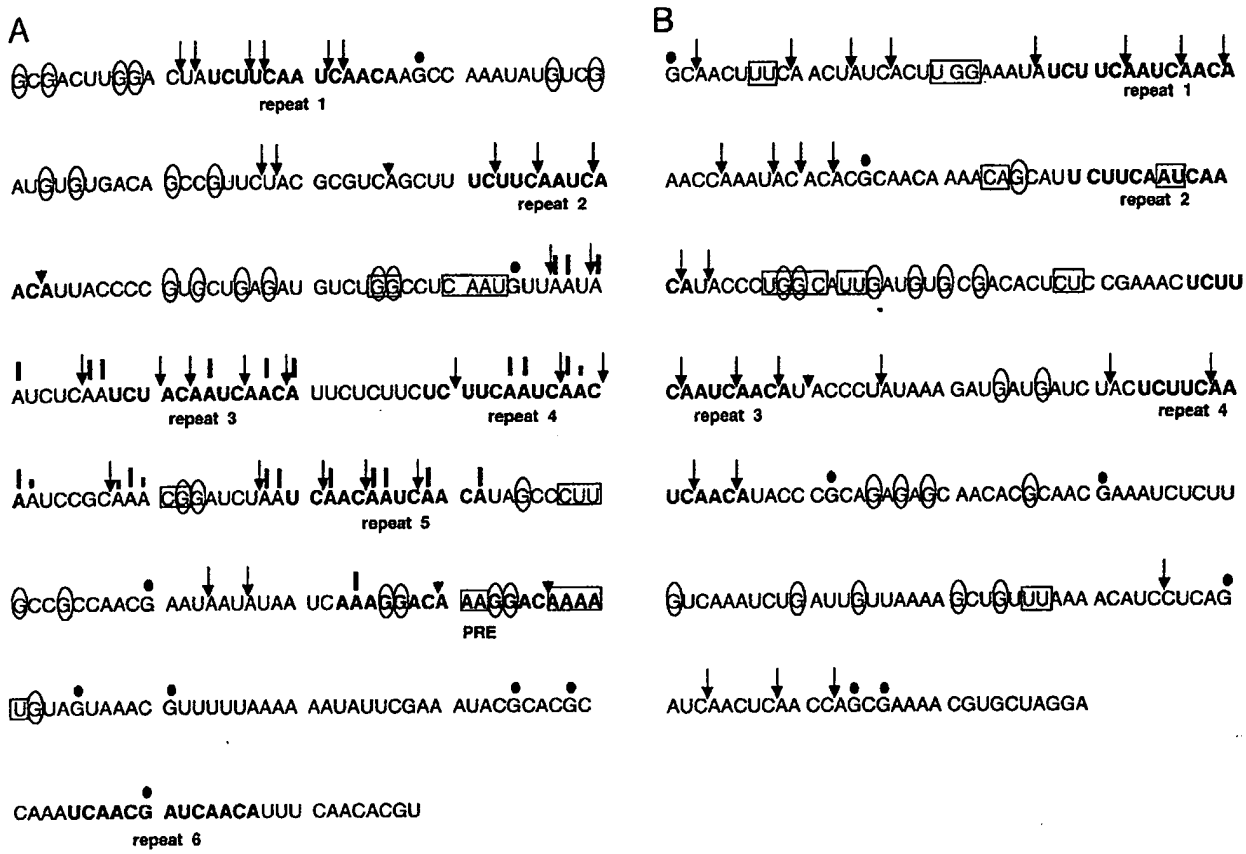


FIGURE 5. Summary of the enzymatic and chemical structure probing of the *dsx*RE from (**A**) *D. melanogaster* and from (**B**) *D. virilis*. Symbols: long arrows, strong RNase A cut; small arrows, weak RNase A cut; boxed nucleotides, moderate to strong RNase V1 cut; solid dots, strong T1 cuts; open ovals, G residues protected from T1 digestion; long vertical bars, strong sites of DMS modification; small vertical bars, weak sites of DMS modification. Repeat elements in each *dsx*RE are in bold.

filters (Gott et al., 1993). Analysis of the oligonucleotides recovered from the filters indicated that retention by Tra binding is very efficient for all truncated RNAs, even for those that contain only *dsx* unrelated polylinker sequences. In contrast, only those RNAs containing one or more complete 13-nt repeat elements were retained on the filter by Tra2 (data not shown). Thus, consistent with previous filter-binding and UV-crosslinking studies (Lynch & Maniatis, 1995, 1996), Tra interacts with RNA with little or no specificity, whereas Tra2 binds specifically to the repeat elements.

To determine whether Tra2 can protect specific regions of the *dsx*RE from RNAse digestion, the R2-5PRE was subjected to RNase A digestion in the presence of increasing concentrations of Tra2. As shown in Figure 7, all of the repeat elements present in R2-5PRE remain accessible to RNase A digestion at Tra2 concentrations lower than 20 nM. Surprisingly, all of the repeat elements except repeat 2 are protected to similar degrees at concentrations of ≥100 nM Tra2. This observation correlates well with the observed binding affinity of $K_d = 50$ nM for Tra2 measured under identical conditions by a nitrocellulose filter-binding assay (data not shown). The absence of selective binding of Tra2 to individual repeats indicates that the 13-nt repeat elements represent multiple Tra2 binding sites within the enhancer region. These binding sites, except repeat 2, are occupied at similar Tra2 concentrations. Similarly, Tra2 binds to the PRE at approximately the same concentration. Interestingly, our RNA structural probing data (Fig. 4) and computer-assisted folding analysis predict that the PRE is primarily in a double-stranded configuration. However, with increasing concentrations of Tra2, part of the PRE sequence becomes more susceptible to RNase A cleavage and part of the region is protected from this cleavage (Fig. 7). Thus, Tra2 binding appears to induce a conformational change in the PRE RNA.

A series of modification interference experiments were performed to identify whether specific nucleotides within the enhancer region are essential for Tra or Tra2 binding. The protocol for modification interference requires conditions to achieve a single modification per RNA molecule and relies subsequently on the

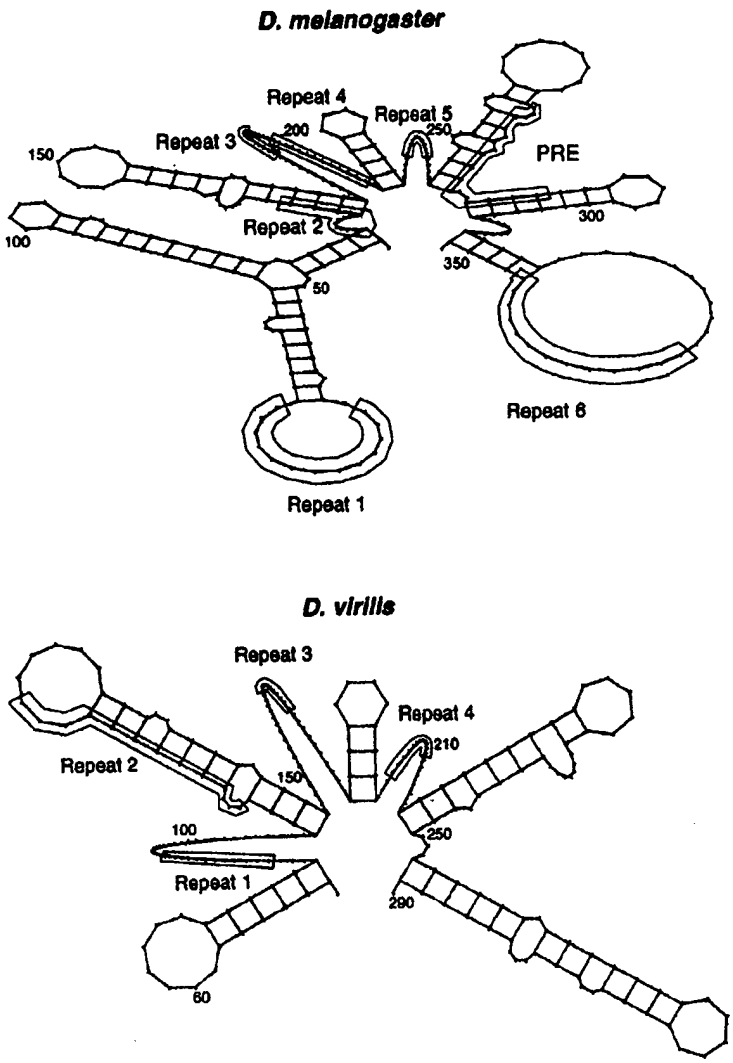


FIGURE 6. Secondary structure models of the *dsx*REs from *D. melanogaster* and from *D. virilis*. The structures were generated as described in the text. Boxed areas indicate the position of the 13-nt repeat elements and the PRE. Standard Watson–Crick base pairs and G·U pairs are indicated by bars.

separation of free RNA from the bound species by a nitrocellulose filter-binding assay (Conway & Wickens, 1989). The comparison of RNA molecules selected by Tra or Tra2 binding with the initial RNA pool did not lead to the identification of nucleotides essential for the interaction of Tra or Tra2 with the enhancer (data not shown). This data supports the above conclusion that the enhancer region contains not one high-affinity site for the interaction with Tra2, but several. Although Tra2 is capable of binding to the *dsx*RE in the absence of additional proteins, we note that the specificity of Tra2 dramatically increases in the presence of SR proteins (Lynch & Maniatis, 1995) and in nuclear extracts (Lynch & Maniatis, 1996) under splicing conditions. Thus, the experiments presented here show only that each of the repeats can bind to Tra2 specifically. They do not address the nature of the RNA–protein complex assembled on a functional *dsx*RE.

Antisense inhibition of *dsx* pre-mRNA splicing

To determine whether the single-stranded configuration of the 13-nt repeats in the *D. melanogaster dsx*RE is required for enhancer function, we conducted experiments to determine whether the repeats can function in a double-stranded configuration. An antisense oligonucleotide complementary to the 13-nt repeat was annealed to the *dsx*RE and the effects on pre-mRNA splicing were examined. As shown in Figure 8, the presence of the antisense oligonucleotide has a dramatic effect on the splicing efficiency of substrate D2. In the absence of antisense oligonucleotide, approximately 25% of the substrate is spliced. In contrast, no spliced products are detected above the degradation levels regardless of whether the substrate was preincubated with the antisense oligonucleotide at conditions identical to those used in the secondary structure anal-

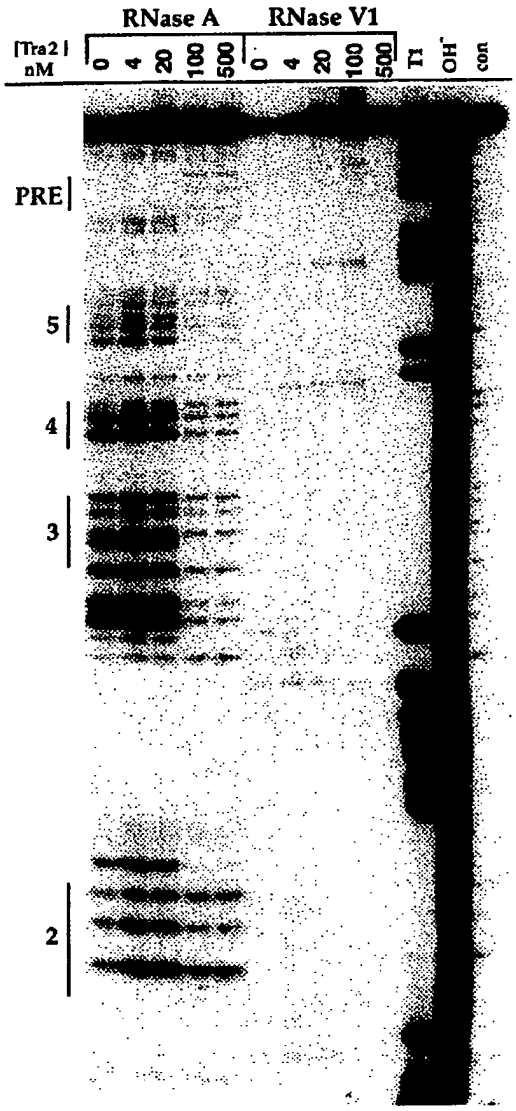


FIGURE 7. RNase protection pattern of R2-5PRE as a function of increasing concentrations of Tra2. Lanes: T1, RNase T1 digestion at denaturing conditions; OH^-, alkaline hydrolysis; con, no RNase added. Eight microliters of 5′-labeled R2-5PRE in reaction buffer was allowed to equilibrate with increasing concentrations of Tra2 at 30 °C for 20 min. Each binding reaction was then adjusted with either 2 μL of a 0.02 U/mL solution of RNase A for 8 min or with 2 μL of a freshly prepared 3.5 U/mL solution of RNase V1 for 10 min at 30 °C. Positions of the 13-nt repeat elements and the PRE are indicated on the left side.

ysis (Fig. 8A), heat annealed prior to the addition of nuclear extract (Fig. 8B), or whether the antisense oligonucleotide was added shortly after the incubation of D2 in nuclear extract (Fig. 8C). Thus, the *dsx* enhancer presents the repeats in a single-stranded configuration in the absence (Fig. 8A) and in the presence of nuclear extract (Fig. 8C). Similar results were obtained in antisense experiments using D1 as the substrate with the exception of a significantly reduced splicing efficiency for heat-treated D1 in the absence of antisense (data not shown). In control experiments, the presence of the antisense oligonucleotide at concentrations that inhibited *dsx* pre-mRNA splicing did not affect the splicing of β-globin pre-mRNA (data not shown). We conclude that the single-stranded character of the 13-nt repeats is essential for splicing enhancer activity. Thus, the conservation of both the sequence and the structure of the 13-nt repeats are essential for *dsx* splicing enhancer activity.

DISCUSSION

On the basis of results presented here, we propose that the *dsx*RE adopts a secondary structure that optimizes interactions between individual repeat elements and the RNA-binding domains of Tra2 and SR proteins. In all cases, except repeat 2, the repeats are present in a single-stranded configuration. This structural difference correlates with the observation that repeat 2 is not protected as efficiently from RNase A digestion at Tra2 concentrations that maximally protect the other repeats. Thus, the single-stranded character of the repeats may be essential for the formation of a stable multiprotein complex consisting of Tra2, Tra, SR proteins, and possibly other nuclear proteins (Tian & Maniatis, 1993). This stable enhancer complex can then facilitate the recruitment of general splicing factors to the upstream female-specific 3′ splice site (Zuo & Maniatis, 1996).

A comparison of the *dsx*REs from *D. melanogaster* and *D. virilis* revealed conserved sequences at the 5′ and 3′ ends of the element, and conserved repeat sequences that are separated by highly divergent RNA sequences. Although attempts to identify a function for the conserved 5′ and 3′ sequences by in vitro splicing assays have thus far failed (K.W. Lynch & T. Maniatis, unpubl.), it seems likely that this conservation is important in flies. The conservation of the repeat sequences is almost certainly due to a conserved recognition by Tra and Tra2, because we have shown that the *D. melanogaster* and *D. virilis dsx*REs are functionally interchangeable, and both require *D. melanogaster* Tra and Tra2 for their functions. In addition, a recent study demonstrated that the *D. virilis* Tra homologue can partially rescue the Tra mutant phenotype in transgenic *D. melanogaster* (O'Neil & Belote, 1992). Although the primary sequence and the detailed secondary structures of the inter-repeat sequences are poorly conserved, our data indicate that the repeats in both species are maintained as single-stranded RNA regions. Thus, the inter-repeat structure may have evolved to maintain the repeats in this configuration. The importance of maintaining the repeats in a single-stranded configuration was demonstrated by showing that a 13-nt RNA complementary to the repeats inhibits enhancer-dependent splicing. Thus, the repeat ele-

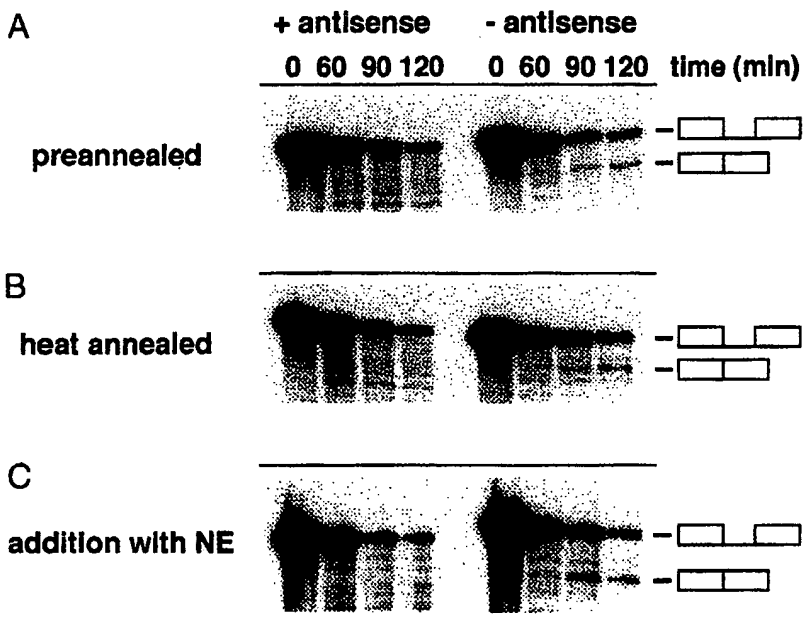


FIGURE 8. Antisense inhibition of *dsx* pre-mRNA splicing. The splicing efficiency of D2 was followed over 120 min in the presence or absence of 4 μM antisense oligonucleotide complementary to the 13-nt repeat element. D2 was (A) preincubated with the antisense oligonucleotide under conditions identical to those used in the secondary structure determination prior to the addition of nuclear extract, (B) heat annealed at splicing conditions to the antisense oligonucleotide prior to the addition of nuclear extract, or (C) incubated in nuclear extract for 1–2 min prior to the addition of antisense oligonucleotide.

ments within the *dsx* enhancer are only recognizable by Tra/Tra2 and SR proteins when presented in a single-stranded configuration.

Comparison of the *D. melanogaster* and *D. virilis* sequences revealed no conservation of the PRE identified in the *dsx*RE from *D. melanogaster* (Lynch & Maniatis, 1995). However, a distinct purine-rich sequence is found in the *D. virilis dsx*RE, and this sequence is capable of functioning as a constitutive splicing enhancer (K.W. Lynch & T. Maniatis, unpubl.). Surprisingly, our RNA structural probing data and the computer-assisted folding analysis indicate that the PRE in *D. melanogaster* is primarily in a duplex configuration, even though Tra, Tra2, and SR proteins crosslink specifically to this sequence in nuclear extracts (Lynch & Maniatis, 1996). RNA footprinting data (Fig. 7) suggests that Tra2 binding induces a conformational change in the PRE.

Although the secondary structure representations of the *dsx*RE, which are shown in Figure 6, are consistent with the biochemical data, it is important to note that each provides only one of many possible and very similar configurations. There are several architectural similarities in the *dsx*RE structures obtained. Repeats 3 and 4 in *D. virilis* are flanked by extended hairpins and separated by a four-base pair hairpin. In *D. melanogaster*, the structural format resembles that observed in *D. virilis*, with the exception that an additional 13-nt repeat element is included. Thus, repeats 3 and 5 are flanked by extended helical regions and repeats 4 and 5 are separated by a small but stable hairpin. In both proposed structures, repeat 2 is involved in the formation of the hairpin that flanks repeat 3.

There is little information to speculate on the functional importance of the proposed secondary structure elements flanking each repeat element. Previously, a single repeat element (Hoshijima et al., 1991) or a synthetic tandem repeat element was shown to substitute for the *D. melanogaster dsx*RE in in vivo transfection experiments (Inoue et al., 1992). Similar results have been obtained in vitro using HeLa cell nuclear extracts (K.J. Hertel & T. Maniatis, in prep.). These observations indicate that the *cis*-regulatory element requirement is met by the presence of only one or two 13-nt repeat elements in close proximity to each other. Thus, the secondary structure elements in the *dsx*RE are not required for Tra- and Tra2-dependent stimulation of splicing as long as the single-stranded character of the repeats is maintained. However, the evolutionary conservation of multiple repeats and their single-stranded nature argue strongly that both are required for the fine-tuned regulation of *dsx* pre-mRNA splicing in flies. For example, five of six or three of four repeat elements in *D. melanogaster* and in *D. virilis*, respectively, are in single-stranded configuration. Thus, the inter-repeat structure elements might influence indirectly the efficiency of splice site activation by maintaining the accessibility of the repeat elements. In flies, where the levels of Tra and Tra2 are likely to be less than those generated in transfection or in vitro experiments, this arrangement may be essential for the controlled function of the *dsx*RE. By contrast, the high levels of Tra and Tra2 used in in vitro experiments or produced in cotransfection experiments are sufficient to observe splice site activation with only a single repeat element.

In addition to the structural analysis of the *dsx*RE, the footprinting and terminal truncation results have shown that the 13-nt repeat elements can act as binding sites for Tra2. The data are therefore consistent with the model of a direct interaction of Tra2 with the 13-nt repeat element. Other lines of evidence indicate

that specific protein–RNA interactions within the repeats are highly dependent on protein–protein interactions. A recent in vitro binding analysis demonstrated cooperative binding of Tra, SR proteins, and Tra2 to the intact *dsx*RE (Lynch & Maniatis, 1995). When assayed in nuclear extracts, efficient binding of Tra2 to the 13-nt repeat is highly dependent on the presence of Tra (Lynch & Maniatis, 1996). Given these observations, it is very likely that the number of repeat sequences in the *dsx*RE and their context-dependent accessibility to Tra, Tra2, and SR proteins results in protein–RNA interactions that lead to the formation of an enhancer complex capable of promoting 3′ splice site recognition at a distance. This property is not shared with simple constitutive enhancer elements.

Tra2 and SR proteins, but not Tra, contain the RRM RNA-binding domain found in a large family of RNA-binding proteins involved in RNA metabolism. The crystal structure of the U1A–snRNA complex showed that the RRM makes specific contacts with the single-stranded region of an RNA hairpin structure (Oubridge et al., 1994). Similarly, the single-stranded nature of the repeat elements of the *dsx*RE is consistent with the possibility that they are recognized by the RRM of Tra2 and possibly SR proteins.

MATERIALS AND METHODS

RNA

D1 RNA was synthesized from plasmid D1 using T7 RNA polymerase (Tian & Maniatis, 1992). The splicing substrate, D2, in which the enhancer region of *D. melanogaster* was substituted with the enhancer region of *D. virilis*, was constructed by subcloning the enhancer region (inclusive of repeat 1 to just before the 3′ conserved region) into a PCR-generated *Eco*R I site of D1 located just upstream of the first repeat sequence (Lynch & Maniatis, 1995).

The probes R2-5PRE, R2-6, and D6 were generated as described previously (Lynch & Maniatis, 1995). The construct encoding R1-6 was made by cloning a fragment containing the T7 transcription start site and repeat 1 into the *Mlu* I site of R2-6; RNA was then synthesized by in vitro transcription. V1-4 was transcribed from a construct in which the *D. virilis* enhancer PCR fragment from D2 was cloned into the *Eco*R I site downstream of the T7 promoter in SP72.

Splicing substrates were labeled uniformly with [^{32}P]UTP. 5′-End-labeling of the oligonucleotides R2-5PRE, R2-6, R1-6, and V1-4 synthesized by T7 RNA polymerase was accomplished by removing 5′ triphosphates with calf intestine phosphatase followed by reaction with [γ-^{32}P]ATP and T4 polynucleotide kinase. Oligonucleotide concentrations were determined from specific activities for radiolabeled RNAs, assuming a residue extinction coefficient of $8.5 \times 10^3\ M^{-1}\ cm^{-1}$ at 260 nm for nonradioactive RNA.

In vitro splicing reactions

In vitro splicing reactions were generally conducted as described in Tian and Maniatis (1992). The RNA antisense experiments were conducted using identical conditions except for the presence of 4 μM antisense oligonucleotide and approximately 50 nM poly I-C. The presence of poly I-C was required to reduce the activity of a double-stranded deaminase (dsRad) activity present in HeLa cell nuclear extracts (Yang et al., 1995). Control experiments with *dsx* pre-mRNA and β-globin pre-mRNA established that the presence of poly I-C did not affect splicing efficiency significantly. The splicing substrate was incubated under splicing conditions with the antisense RNA oligonucleotide for 30 min at 30 °C with or without a prior heat anneal step (95 °C for 1.5 min in the absence of $MgCl_2$, then add $MgCl_2$). The splicing reaction was then initiated by the addition of nuclear extract, poly I-C, and Tra/Tra2. The final concentrations were 30% (v/v) nuclear extract, 4 μM antisense RNA, 50 nM poly I-C, 50 nM Tra, and 50 nM Tra2 in a volume of 50 μL. In another experiment, the substrate was incubated with nuclear extract for 1–2 min prior to the addition of poly I-C, antisense RNA, and Tra/Tra2.

Recombinant proteins

Recombinant Tra and Tra2 were expressed in baculovirus and purified as described in Tian and Maniatis (1992).

Cloning and sequencing of the *D. virilis doublesex* female-specific exon

The female-specific exon of *D. virilis doublesex* gene was isolated from a *D. virilis* genomic library constructed in EMBL3, which was kindly provided by Stuart Newfeld. The library was screened through successive rounds of high-stringency hybridization to a DNA probe that contained the sequence of the third intron of the *D. virilis doublesex* gene. The probe was isolated by PCR from *D. virilis* genomic DNA using primers based on the published *D. virilis doublesex* intron sequence (Burtis & Baker, 1989). After a positive clone was identified, the phage DNA was isolated, digested with various restriction enzymes, and analyzed by Southern blot to determine the minimal fragment that contained the intron sequence. A 1.8-kDa *Bst*Y I fragment, which hybridized strongly to the intron probe, was then subcloned into SP73 and sequenced using the T7 and SP6 priming sites.

Enzymatic and chemical structure probing

All reactions were conducted at 30 °C in a buffer containing 72 mM KCl, 12 mM HEPES, pH 7.9, 3.2 mM $MgCl_2$, 1 mM ATP, 20 mM creatine phosphate, and 4% glycerol. These conditions were chosen to mimic those used in the splicing reaction. For the enzymatic probing of the enhancer RNAs, the RNases A, T1, and V1 were used. RNase A and T1 are single-stranded and nucleotide-specific RNases leaving 3′-phosphate products. RNase A cleavage is pyrimidine-specific with a preference for CpN bonds (Knapp, 1989). RNase T1 recognizes GpN bonds. Both enzymes remain active in EDTA. RNase V1 was used to determine which portions of the RNA are found base paired at the conditions used. V1 requires the presence of Mg^{2+} for activity. In a typical structure probing experiment, a trace amount of 5′ end-labeled RNA in 10 μL reaction buffer was incubated with either 0.02 U/mL RNase A, 0.4 U/mL T1, or 0.07 U/mL V1.

Time points (3 μL) were taken at appropriate time intervals, mixed with a formamide buffer containing 10 mM EDTA, 0.02% bromophenol blue, and 0.02% xylene cyanol and immediately frozen to −70 °C until all time points were collected. Each time point was then subjected to 6% PAGE. For all RNA probes tested, the presence of carrier tRNA or the addition of a denaturing/renaturing step prior to the digestion did not result in an altered susceptibility to the RNases used.

In addition to the enzymatic probing, chemical base modification assayed by reverse transcription was used to examine the secondary structure of the *D. melanogaster* enhancer region. The DMS treatment was conducted according to Zaug and Cech (1995). The modified RNAs were annealed to a 20-nt [γ-^{32}P]-end-labeled DNA primer, ATTTTGTCCTTTGT CCTTTG, corresponding to sequences between the fifth and sixth repeats. The annealed primer/RNA complex was then extended with SuperScript reverse transcriptase (Gibco BRL). Typically, 0.5 μg of R2-5PRE in 50 μL of reaction buffer (90 nM R2-5PRE) was incubated for varying times with 1-3 μL of a 30% (v/v) DMS/ethanol mix. Each time point was quenched with 0.5× volume of 0.75 M sodium acetate and 0.5 M β-mercaptoethanol. After ethanol precipitation and resuspension, approximately 0.1 μg of the modified RNA in 10 μL (80 nM R2-5PRE) was annealed to a fourfold molar excess of 5′-end-labeled DNA primer. Each of the 10-μL extension reactions used 2 μL of the annealed primer mixture and were supplemented with 450 mM dNTPs and 100 U of SuperScript reverse transcriptase. After 1 h at 42 °C, reactions were terminated by the addition of an equal volume of formamide buffer, heated to 95 °C for 2 min, and then subjected to 6% PAGE.

Computer analysis

The primary sequence data from the enhancer regions of *D. melanogaster* and *D. virilis* were aligned by computer to determine primary sequence conservation. They were then analyzed independently with the MFOLD and PLOTFOLD application programs of the Sequence Analysis Software Package (version 7.2) from the Genetics Computer Group, University of Wisconsin Biotechnology Center. Version 7.2 makes use of an RNA-folding algorithm developed by Zuker (Jaeger et al., 1989) and incorporates updated bond energies. The secondary structure data accumulated in experiments described above was used to restrict the folding of nucleotides that are predominantly in a single-strand conformation. MFOLD generates optimal and suboptimal structures. These were then analyzed for agreement with the remaining experimental data.

ACKNOWLEDGMENTS

We thank Stuart Newfeld for providing the *D. virilis* genomic library and members of our laboratory for discussion and comments on the manuscript. This investigation was supported by a postdoctoral fellowship from the Jane Coffin Childs Memorial Fund for Medical Research (K.J.H.) and grant GM 42231 from the National Institutes of Health (T.M.).

Received May 13, 1996; returned for revision June 21, 1996; revised manuscript received August 2, 1996

REFERENCES

Amrein H, Gorman M, Nöthiger R. 1988. The sex-determining gene *tra*-2 of *Drosophila* encodes a putative RNA binding protein. *Cell* 55:1025–1035.

Amrein H, Hedley ML, Maniatis T. 1994. The role of specific protein-RNA and protein-protein interactions in positive and negative control of pre-mRNA splicing by *Transformer* 2. *Cell* 76:735–746.

Baker BS. 1989. Sex in flies: The splice of life. *Nature* 340:521–524.

Bandziulis RJ, Swanson MS, Dreyfuss G. 1989. RNA-binding proteins as developmental regulators. *Genes & Dev* 3:431–437.

Blackman RK, Meselson M. 1986. Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila* hsp82 gene. *J Mol Biol* 188:499–515.

Boggs RT, Gregor P, Idriss S, Belote JM, McKeown M. 1987. Regulation of sexual differentiation in *D. melanogaster* via alternative splicing of RNA from the *transformer* gene. *Cell* 50:739–747.

Burtis KC, Baker BS. 1989. *Drosophila doublesex* gene controls somatic sexual differentiation by producing alternative spliced mRNAs encoding related sex-specific polypeptides. *Cell* 56:997–1010.

Conway L, Wickens M. 1989. Modification interference analysis of reactions using RNA substrates. *Methods Enzymol* 180:369–379.

Dominski Z, Kole R. 1994. Identification of exon sequences involved in splice site selection. *J Biol Chem* 269:23590–23596.

Fu XD. 1995. The superfamily of arginine/serine-rich splicing factors. *RNA* 1:663–680.

Gott JM, Pan T, LeCuyer KA, Uhlenbeck OC. 1993. Using circular permutation analysis to redefine the R17 coat protein binding site. *Biochemistry* 32:13399–13404.

Hedley ML, Maniatis T. 1991. Sex-specific splicing and polyadenylation of *dsx* pre-mRNA requires a sequence that binds specifically to *tra*-2 protein in vitro. *Cell* 65:579–586.

Heinrichs V, Baker BS. 1995. The *Drosophila* SR protein RBP1 contributes to the regulation of *doublesex* alternative splicing by recognizing RBP1 RNA target sequences. *EMBO J* 16:3987–4000.

Hoshijima K, Inoue K, Higuchi I, Sakamoto H, Shimura Y. 1991. Control of *doublesex* alternative splicing by *transformer* and *transformer-2* in *Drosophila*. *Science* 252:833–836.

Inoue K, Hoshijima K, Higuchi I, Sakamoto H, Shimura Y. 1992. Binding of the *Drosophila* Transformer and Transformer-2 proteins to the regulatory elements of *doublesex* primary transcripts for sex-specific RNA processing. *Proc Natl Acad Sci USA* 89:8092–8096.

Jaeger JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA* 86:7706–7710.

Knapp G. 1989. Enzymatic approaches to probing of RNA secondary and tertiary structure. *Methods Enzymol* 180:192–212.

Krol A, Carbon P. 1989. A guide for probing native small nuclear RNA and ribonucleoprotein structures. *Methods Enzymol* 180:212–227.

Lynch KW, Maniatis T. 1995. Synergistic interactions between two distinct elements of a regulated splicing enhancer. *Genes & Dev* 9:284–293.

Lynch KW, Maniatis T. 1996. Assembly of specific SR protein complexes on distinct regulatory elements of the *Drosophila doublesex* splicing enhancer. *Genes & Dev* 10:2089–2101.

Maniatis T. 1991. Mechanisms of alternative pre-mRNA splicing. *Science* 251:33–34.

Nagai K, Oubridge C, Ito N, Avis J, Evans P. 1995. The RNP domain: A sequence-specific RNA binding domain involved in processing and transport of RNA. *Trends Biochem Sci* 20:235–240.

Newfeld SJ, Smoller DA, Yedvobnick B. 1991. Interspecific comparison of the unusually repetitive *Drosophila* locus *mastermind*. *J Mol Evol* 32:415–420.

Noller HF. 1984. Structure of ribosomal RNA. *Annu Rev Biochem* 53:119–162.

O'Neil MT, Belote JM. 1992. Interspecific comparison of the *transformer* gene of *Drosophila* reveals an unusually high degree of evolutionary divergence. *Genetics* 131:113–128.

Oubridge C, Ito N, Evans PR, Teo C-H, Nagai K. 1994. Crystal structure at 1.92 Å resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* 372:432–438.

Ramchatesingh J, Zahler AM, Neugebauer KM, Roth MB, Cooper TA. 1995. A subset of SR proteins activates splicing of the cardiac troponin T alternative exon by direct interactions with an exonic enhancer. *Mol Cell Biol* 9:4898–4907.

Ryner LC, Baker BS. 1991. Regulation of doublesex pre-mRNA pro-

cessing occurs by 3′-splice site activation. *Genes & Dev* 5:2071–2085.

Sun Q, Mayeda A, Hampson RK, Krainer AR, Rottmann FM. 1993. General splicing factors SF2/ASF promote alternative splicing by binding to an exonic splicing enhancer. *Genes & Dev* 7:2598–2608.

Tian M, Maniatis T. 1992. Positive control of pre-mRNA splicing in vitro. *Science* 256:237–240.

Tian M, Maniatis T. 1993. A splicing enhancer complex controls alternative splicing of *doublesex* pre-mRNA. *Cell* 74:105–114.

Tian M, Maniatis T. 1994. A splicing enhancer exhibits both constitutive and regulated activities. *Genes & Dev* 8:1703–1712.

van Oers CCM, Adema GJ, Zandberg H, Moen TC, Bass PD. 1994. Two different sequence elements within exon 4 are necessary for calcitonin-specific splicing of the human calcitonin/calcitonin gene-related peptide I pre-mRNA. *Mol Cell Biol* 14:951–960.

Watakabe A, Tanaka K, Shimura Y. 1993. The role of exon sequences in splice site selection. *Genes & Dev* 7:407–418.

Wu JY, Maniatis T. 1993. Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* 75:1061–1070.

Yang JH, Sklar P, Axel R, Maniatis T. 1995. Editing of glutamate receptor subunit B pre-mRNA in vitro by site specific deamination. *Nature* 374:77–81.

Zaug AJ, Cech TR. 1995. Analysis of the structure of *Tetrahymena* nuclear RNAs in vivo: Telomerase RNA, the self-splicing rRNA intron and U2 snRNA. *RNA* 1:363–374.

Zuo P, Maniatis T. 1996. The splicing factor U2AF^{35} mediates critical protein–protein interactions in constitutive and enhancer-dependent splicing. *Genes & Dev* 10:1356–1368.

ELSEVIER

Insect Biochemistry and Molecular Biology 34 (2004) 149–157

Insect Biochemistry and Molecular Biology

www.elsevier.com/locate/ibmb

Medfly promoters relevant to the sterile insect technique

Katia Komitopoulou [a], George K. Christophides [a], Katerina Kalosaka [b], George Chrysanthis [b], Maria A. Theodoraki [b], Charalambos Savakis [c], Antigone Zacharopoulou [b], Anastassios C. Mintzas [b,*]

[a] *Department of Genetics and Biotechnology, School of Biological Sciences, University of Athens, Greece*
[b] *Division of Genetics, Cell and Developmental Biology, Department of Biology, University of Patras, Greece*
[c] *Institute for Molecular Biology and Biotechnology-Foundation for Research and Technology, Hellas, Greece*

Received 2 December 2002; received in revised form 4 April 2003; accepted 12 June 2003

Abstract

This review summarizes structural and functional studies on medfly promoters and regulatory elements that can be used for driving sex-specific, conditional and constitutive gene expression in this species. Sex-specific and conditional promoters are important for generating transgenic sexing strains that could increase the performance of the Sterile Insect Technique while strong constitutive promoters are necessary for developing sensitive transgenic marker systems. The review focuses on the functional analysis of the promoters of two male-specific and heat shock medfly genes. A special emphasis is put on the potential utility of these promoters for developing transgenic sexing strains.

Keywords: Transgenic insects; Sex-specific promoters; Heat shock promoters; *Ceratitis capitata*; *Drosophila melanogaster*; Genetic sexing strains; Sterile insect technique

1. Introduction

Fruit flies in the family Tephritidae are rated among the world's most destructive agricultural pests, especially in commercial fruit and vegetables. Chemical pesticide control is the most commonly used method for containing fruit fly populations with known adverse effects on the environment and health. During the last decades, there has been an increasing interest in biological methods for control of insect pests aiming at replacing the existing insecticide-based control methods. A biological method that has proven to be effective in the field for the area-wide control of some insects is the sterile insect technique (SIT). SIT is a species-specific and environmentally non-polluting method of insect control that relies on the mass rearing, sterilization, and release of a large numbers of insects (Knipling, 1955; Krafsur, 1998). If enough sterile insects are released for a sufficient time, most of the wild females in the field mate with the released sterile males and thus produce no viable offspring. Highly successful, area-wide SIT programs have been operated against major agricultural pests such as the New World screwworm, *Cochliomyia hominovorrax*, the tsetse fly (*Glossina* spp.) and the Mediterranean fruit fly (medfly) *Ceratitis capitata* (reviewed in Robinson, 2002).

The medfly is a notorious pest with a worldwide range and a history of fast expansion and painful invasions to various countries (Harris, 1989), and so far it is the best-studied fruit fly at the genetic and molecular level. For medfly, SIT has been shown to be most effective when only sterile males are released in the field (Hendrichs et al., 1995). Current medfly SIT programs use genetic sexing strains (GSS) that are based on the use of male linked chromosomal translocations, where the translocation carries a dominant wild-type allele for a selectable gene. These chromosome aberration-based systems tend to be unstable and reduce the fitness of the insects, making them less effective agents for SIT (Robinson et al., 1999). In addition, analogous strains

* Corresponding author. Tel.: +30-610-997-368; fax: +30-610-997-881.
E-mail address: mintzas@upatras.gr (A.C. Mintzas).

0965-1748/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2003.06.016

Exhibit F

have to be constructed, *de novo*, for each target species, a laborious task for insects with a limited genetic background.

An alternative method for making GSS is to use genetic engineering (Alphey and Andreasen, 2002) as has been recently demonstrated in *Drosophila melanogaster* (Thomas et al., 2000; Heinrich and Scott, 2000; Markaki et al., this issue). In the past six years, stable genetic transformation systems developed in several pest insects, including medfly, provided significant opportunities to further improve the effectiveness of SIT and to develop novel pest control strategies (reviewed in Atkinson et al., 2001). Gene transfer technology can lead to two major improvements of SIT: (a) development of transgenic sexing systems for the generation of novel GSS with better characteristics than the existing ones and (b) development of transgenic marker systems for detecting, maintaining and recognizing transgenic insects. A major advantage of these systems is that they are likely to be applied in a wide range of pest insects. Sex-specific and conditional promoters and regulatory elements are key components for developing transgenic sexing systems, while strong constitutive promoters are important for developing sensitive transgenic marker systems. A number of promoters from sex-specific, conditional and constitutive medfly genes have been cloned. In the present review we summarize published data on the structural and functional characterization of these promoters and present new data on the functional analysis of a conditional *hsp*70 promoter.

2. Male-specific promoters

Transgenic sexing strains can be constructed by using male-specific promoters to drive the expression of selectable genes encoding 'resistance factors' in males. These strains could be grown under normal conditions, and then switched to restrictive conditions for the last generation so that all females die, giving a male only population for sterile release programs.

Five male-specific serum proteins (MSSPs) have been characterized in the medfly (Katsoris et al., 1990; Thymianou et al., 1995). The two major ones are homodimers of two related polypeptides (MSSP-α and -β), with molecular weights of 14.5 and 13.5 kDa respectively, while the others are homo- and hetero-dimers of α- and β-type polypeptides. By screening an expression library with anti-MSSP antibodies, a cDNA coding for an α-type polypeptide with structural similarities to the odorant binding proteins was isolated (Thymianou et al., 1998). A small multigene family encoding closely related MSSP polypeptides was subsequently cloned and characterized. This family consists of at least seven members, divided according to sequence similarity in three subgroups, two closely related *MSSP*-α (α*1*, α*2*) and *MSSP*-β (β*1*, β*2*, β*3*), and one more divergent, *MSSP*-γ (γ*1*, γ*2*) (Christophides et al., 2000a). Phylogenetic analysis of the MSSP gene family showed that it has originated by gene duplications of an ancestral gene. The very high degree of identity, both in their coding and surrounding regions, predicts that MSSP genes have arisen by very recent gene duplications. Although MSSPs are mainly expressed in the male fat body, analytical expression studies by RNA blot hybridization and RT-PCR suggested that individual members of this family are expressed in a distinct sex- and tissue-specific manner (Christophides et al., 2000a).

2.1. Functional analysis of *MSSP*-α2 and *MSSP*-β2 promoters

The *MSSP-α2* and *MSSP-β2* genes have identical 5′ untranslated regions (5′ UTR) and exhibit 94.5% identity along their 504 bp upstream promoter regions, presenting a few nucleotide substitutions and single or small nucleotide deletions (Christophides et al., 2000b). In both genes, a putative transcription initiation site is located 37 bp upstream of the ATG initiation codon and 31 bp downstream of a typical TATA box. Functional analysis of the promoters of these genes was performed in transgenic medflies using the *Minos* transformation system (Christophides et al., 2000b). For the construction of the *Minos*-based transposon plasmids presented in Fig. 1A, two overlapping promoter fragments of each gene containing the 5′ UTRs and additional 5′ flanking regions were fused to the recombinant AUGβ-gal (*lacZ*) reporter gene (Mismer and Rubin, 1987; Thummel et al., 1988). As shown in Fig. 1A, the two overlapping promoter fragments (α2PS and α2PL) correspond to −283/+37 and −522/+37 sequences of the *MSSP-α2* gene and the analogous fragments (β2PS and β2PL) correspond to −287/+37 and −485/+37 sequences of the *MSSP-β2* gene.

The results from the transformation experiments with the four *MSSP-lacZ* constructs are shown in Table 1. Twenty-nine transgenic lines were established from a total of 1557 G0 adults. In all α2PS and α2PL lines, *lacZ* expression was exclusively detected in the fat body of adult males. In α2PS lines, X-gal staining was detected approximately 72 h after eclosion whereas in α2PL lines within the first 30 h after eclosion. Relative to the endogenous MSSP expression, which starts 24 h after eclosion (Thymianou et al., 1995), the expression of the transgene was delayed by 50 h in α2PS lines but correct in α2PL lines. Quantitative measurements of the β-galactosidase activity and western analysis during adult development, in a α2PL line, showed that the expression pattern of the transgenic protein was very similar to that of the endogenous protein. Quantification of β-galactosidase levels in

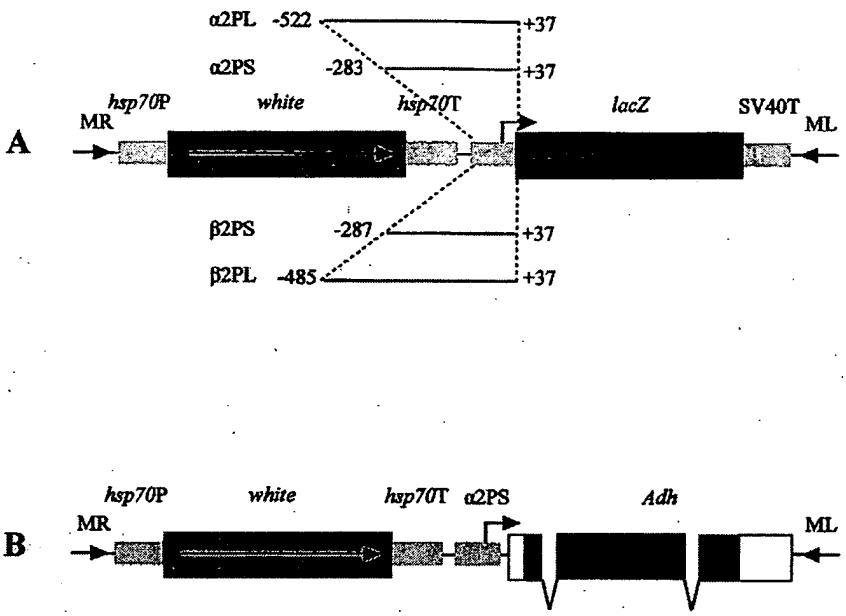


Fig. 1. Schematic illustration of the *Minos* constructs used for promoter analysis of the *MSSP-α2* and *MSSP-β2* genes. (A) Four overlapping fragments from the 5′ region of the genes (α2PS, α2PL, β2PS and β2PL) were fused to a *lacZ* reporter gene and then cloned into the unique *Not* I restriction site of a modified *Minos* transformation vector, pTZMiCcwNotI, which is marked with the medfly *white* gene (Loukeris et al., 1995). *hsp70*P, *hsp70*T and SV40T represent *Drosophila* and SV40 promoter and terminator sequences (Christophides et al., 2000b). (B) The −283/+27 5′ region of the *MSSP-α2* gene (α2PS) was fused to the *Adh gene* encoding the FAST isoform of *Drosophila* ADH and then cloned into the *Minos* transformation vector, pTZMiCcwNotI as described above. The sequence of the *Adh gene* included 36 bp from the 5′UTR, the entire 3′ UTR and 120 bp from the 3′ flanking region is indicated by white boxes.

synchronized transgenic males showed that *lacZ* expression in most α2PL lines was two to twenty-fold higher than that of the strongest α2PS line. Analysis of the β2PS and β2PL lines revealed that the reporter gene was not expressed in the male fat body but in the midgut of both sexes. In both lines, *lacZ* expression started at the pupal stage, a few hours before eclosion, reached maximum levels at the second day, and then declined. Similarly to *MSSP-α2* promoter, the long *MSSP-β2* promoter fragment (β2PL) gave higher β-galactosidase levels than the respective short fragment (β2PS).

In conclusion, these data indicate that the −283/+37 promoter region of the *MSSP-α2* gene is able to drive basal sex-specific gene expression in the fat body of adult males and that additional sequences in the −522/−284 5′ flanking region of the gene are responsible for transcriptional enhancement and correct temporal expression. On the other hand, the −287/+37 promoter region of the *MSSP-β2* gene drives basal expression in the midgut of both sexes. The −485/−288 5′ flanking region of this gene does not affect the sex and tissue specificity but it may confer transcriptional enhancement similar to the −522/−284 region of the *MSSP-α2* gene. Since the *MSSP-α2* and *MSSP-β2* genes have identical 5′ UTRs, the regulatory elements responsible for the differences in the tissue and sex specificity of their promoters must be located in their −283/−1 and −287/−1 regions, respectively. These regions have 12 nucleotide variations, two single deletions and one deletion of 5 nucleotides, all dispersed along their sequence. It seems, therefore, that within the *cis*-acting elements of these regions, a few nucleotides with strong binding affinities for transcription factors may be responsible for the differential function of the *MSSP-α2* and *MSSP-β2* promoters.

The function of *MSSP-α2* and *MSSP-β2* promoters has also been studied in a heterologous system by transforming *D. melanogaster* with α2PL- and β2PL-*lacZ* constructs. Interestingly, both promoters drove transgene expression in the midgut of both *Drosophila* sexes giving identical *lacZ* expression patterns to those obtained in β2PS and β2PL medfly lines. Several

Table 1
Transformation experiments with *MSSP* promoter constructs

Constructs	G0 adults	Transgenic lines	Transformation frequency (%)[a]
α2PS-*lacZ*	404	4	1
α2PL-*lacZ*	368	13	3.5
β2PS-*lacZ*	486	9	2.5
β2PL-*lacZ*	299	3	0.7
α2PS-*Adh*	514	13	2.5

[a] Transformation frequency was calculated as percentage of transgenic lines per G0 adults.

studies have shown that sex-specific expression of genes is poorly conserved between species (reviewed in Schutt and Nothiger, 2000). For example, trypsin genes of *Anopheles gambiae* that are expressed only in adult female mosquitoes are expressed in both sexes of *Drosophila* transformants (Muller et al., 1995; Skavdis et al., 1996). The same phenomenon has been also observed for the apyrase gene of *Anopheles gambiae* which is expressed in the adult salivary glands of female mosquitoes. In *Drosophila* transformants, the apyrase promoter, although maintaining its tissue and temporal pattern, was expressed in both sexes (Lombardo et al., 2000). Since it is known that the genetic basis of sex determination varies widely among insects, these results may reflect the fundamental differences of the regulatory networks that affect sex-specific gene expression among insects.

2.2. A potential sexing system based on male-specific promoters

The idea of using the gene of alcohol dehydrogenase (ADH) for medfly genetic sexing was proposed many years ago (Robinson et al., 1986). In *D. melanogaster*, an ADH-based genetic sexing strain was constructed by combining a translocation of an Adh^+ allele to the Y chromosome with an Adh^- line (Robinson and Van Heemert, 1981). As a first step towards developing an ADH-based genetic sexing system in the medfly, transgenic lines carrying the *Drosophila Adh-F* gene under the direction of the basal α2PS promoter were constructed (Fig. 1B) and tested for alcohol tolerance (Christophides et al., 2001). The results from this transformation experiment are shown in Table 1. Comparison of all the results shown in this table shows that transformation frequencies vary between experiments, most likely due to experimental manipulations. On average, the *Minos* element yields similar transformation frequencies to the *PiggyBac* element in medfly (Handler et al., 1998). The established α2PS-*Adh-F* transgenic lines were designated as MAD (Christophides et al., 2001). Western analysis showed significant amounts of *Drosophila* ADH in adult males of several of these lines. Northern analysis confirmed the male-specific expression of the *Adh* transgene. ADH activity assays showed that both transgenic and endogenous ADHs catalyzed the oxidation of ethanol and 2-propanol. Toxicity tests performed with two MAD lines showed that the difference in the ADH levels between males and females was not enough for achieving genetic sexing, although a slightly increased tolerance was observed in males. However an efficient sexing strain could be made, by using an Adh^- medfly strain as host for transformation and/or the α2PL promoter for driving *Adh* expression. As described above, the activity of this promoter is much higher than the basal α2PS promoter used for constructing the MAD lines. Additionally, elimination or modification of the 3′ UTR negative transcriptional regulatory module (AAGGCTGA) of the *Drosophila Adh* gene (Parsch et al., 1999, 2000) may further increase the ADH levels and the effectiveness of the sexing strain. Using more than one independent insertion would further increase the robustness of the strain, though probably at the cost of some loss of fitness for each extra insertion.

3. Female-specific promoters

Sexing systems, based on engineered conditional female-specific lethal genes, have recently been developed and demonstrated to work efficiently in *D. melanogaster* (Thomas et al., 2000; Heinrich and Scott, 2000; Markaki et al., this issue). In these systems, the regulatory elements of the *Drosophila* major yolk protein (*yp*) genes were used to drive, directly or indirectly, female-specific expression of the conditional lethal genes. Well characterized promoters and upstream regulatory elements of female-specific genes from medfly may be required for developing similar sexing systems in this species. Although a number of medfly female-specific genes have been cloned and characterized, detailed functional analysis of their promoters and regulatory elements has not been conducted. These genes encode for yolk proteins, chorion proteins and antibacterial peptides (ceratotoxins). Data on the structural and functional analysis of these genes are summarized below.

3.1. Yolk protein genes

The two major yolk proteins (Vitellogenins) of the medfly (Vg-1 and Vg-2) are synthesized exclusively in the fat body and the ovaries of the adult females (Rina and Mintzas, 1988). Four vitellogenin genes have been cloned and the sequences of two of them (*vg1-γ* and *vg2-δ*) have been determined (Rina and Savakis, 1991). The 5′ flanking regions of these genes show no significant homology to the respective regions of the *Drosophila yp* genes although several short nucleotide sequences have been conserved between the two species. A number of regulatory elements that are responsible for the sex- and tissue-specific expression of the *Drosophila yp* genes have been well characterized (Garabedian et al., 1986; Logan et al., 1989; Ronaldson and Bownes, 1995). Similar functional studies are necessary for characterizing such elements in the medfly *vg* genes.

3.2. Chorion genes

The chorion genes are expressed in the ovaries of the adult females during the last phase of oogenesis. The

main regulatory elements that are responsible for the sex- and tissue-specific expression of these genes have been characterized in *Drosophila* (Swimmer et al., 1990, 1992; Mariani et al., 1996). Six major chorion genes have been isolated from medfly (Konsolaki et al., 1990; Tolias et al., 1990; Vlachou et al., 1997). Sequence comparisons of four medfly chorion genes with the respective genes of four distantly related *Drosophila* species revealed the presence of well conserved sequences in their 5′ flanking regions that correspond to tissue, temporal and amplification control elements of *D. melanogaster* genes. (Vlachou and Komitopoulou, 2001). Functional studies on the promoter of the medfly *s36* gene in *Drosophila* transformants showed that it operates in a similar manner to the *Drosophila* homolog (Tolias et al., 1993).

3.3. Ceratotoxin genes

Ceratotoxins are closely related antibacterial peptides produced in the female reproductive accessory glands of the medfly (Marchini et al., 1997). Ceratotoxin genes are X-linked and organized in a 26 kb cluster (Rosetto et al., 1997; Rosetto et al., 2000). Ceratotoxin transcripts are detected only in adult females and show maximum levels 6–7 days after eclosion. The presence of highly conserved motifs in the upstream regions of these genes suggests the presence of common regulatory elements. Functional studies are needed to investigate whether these conserved motifs contain important control elements.

4. Conditional and constitutive promoters

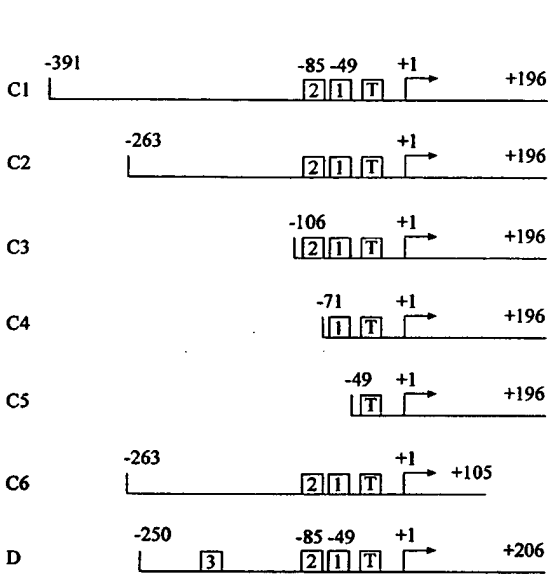
4.1. The hsp70 promoter

Conditional promoters, such as the heat-inducible *hsp70* promoter, could be used for developing sexing systems by driving tansgenes whose expression would lead either to female lethality or to female sex conversion (Pane et al., 2002). The *D. melanogaster hsp*70 promoter has been a popular choice for driving conditional expression of genes in other insect species. A great number of studies have demonstrated the ability of this promoter to function in heterologous systems (Bienz and Pelham, 1982; Voellmy and Rungger, 1982; Burke and Ish-Horowicz, 1982; Mirault et al., 1982; Lis et al., 1982; McMahon et al., 1984; Berger et al., 1985; Atkinson and O'Brochta, 1992). However the activity of this promoter in non-drosophilid insects was found relatively low comparatively to that in *Drosophila* (Berger et al., 1985; Atkinson and O'Brochta, 1992). These data suggest that homologous *hsp*70 promoters should be employed if high gene expression in other insects is required.

Six medfly *hsp*70 genes, organized in two 30 kb coatings, have been isolated (Papadimitriou et al., 1998). All medfly *hsp*70 genes are mapped to the same polytene chromosome band (3L:24C), corresponding to one of the major heat shock puffs. One of these genes (Cc*hsp*70-B1) encodes a 70 kDa protein with 84% amino acid sequence identity to the heat shock 70 proteins of *D. melanogaster*. Similar to the *D. melanogaster hsp*70 genes, the medfly homolog has a long A-rich 5′ UTR and an AT-rich 3′ UTR. These sequences are important for efficient translation under heat shock conditions and for the degradation of the heat shock mRNAs under normal conditions (reviewed in Lindquist and Petersen, 1990). The Cc*hsp*70-B1 gene has two characteristic heat shock elements (HSEs), proximal to the TATA box, that match the heat shock consensus sequence very well, CTnGAAnnTTCnAG (Pelham, 1982) and include three contiguous nGAAn units arranged in alternating orientation characterizing a functional HSE (Amin et al., 1988). These HSEs are located in the region −85/−49, relatively to the putative transcription start site, similarly to the two proximal HSEs of the *Drosophila* 87C1 *hsp*70 gene (Ingolia et al., 1980) which have been shown to be sufficient for optimal expression (Dudler and Travers, 1984; Simon et al., 1985).

4.2. Functional analysis of the hsp70 promoter

Functional analysis of the Cc*hsp*70-B1 promoter was carried out in vivo, in transfected medfly embryos, using the chloramphenicol acetyl transferase-encoding gene (*cat*) as a reporter gene. Six *hsp*70-*cat* constructs, shown in Fig. 2, were made by subcloning PCR amplified overlapping fragments from the 5′ region of the Cc*hsp*70-B1 gene into the pC4cat vector (Thummel et al., 1988). The constructs C1, C2, C3, C4 and C5 contained 391, 263, 106, 71 and 49 bp upstream sequences of the Cc*hsp*70-B1 gene, respectively, and the entire 5′ UTR (196 bp). The construct C6 had the same 5′ flanking region with C1, but contained only the first 105 bp of the 5′ UTR. One *Droshophila hsp*70-*cat* construct (D) was also used for comparison. This construct contained the 456 bp promoter region of the *D. melanogaster* 87C1 *hsp*70 gene (Ingolia et al., 1980) encompassing the three proximal HSEs and the 5′ UTR. Plasmid DNA from the *hsp*70-*cat* constructs was injected into preblastoderm medfly embryos and CAT activity was subsequently measured in embryonic extracts as described by Atkinson and O'Brochta (1992). CAT activity could be detected in 6-h-old embryos, reached maximum levels in 24-h-old embryos and remained relatively constant till the end of embryogenesis (Fig. 3). In 1-day-old larvae no detectable activity was observed. Fig. 2 summarizes the results obtained from five independent experiments for each



	Relative CAT activity	
	Control	Heat shock
C1	77.4±11.1	97.8±13.7
C2	73.2±10.7	100 ±13.5
C3	18.3±5.3	21.5±8.7
C4	2.2±1.2	2.1±1.9
C5	0.5±0.4	0.4± 0.4
C6	75.3±11.8	93.5±10.7
D	56.8±16.4	59.1±18.8

Fig. 2. Functional analysis of the medfly *hsp*70 promoter in transfected medfly embryos. *hsp70-cat* constructs were made by cloning overlapping fragments from the 5′ region of the medfly *hsp*70 gene (C1–C6) and the −250/+206 fragment from the 5′ region of the *D. melanogaster* 87C1 *hsp*70 gene (Ingolia et al., 1980) (D) into the pC4cat vector containing the *cat* gene (Thummel et al., 1988). Boxes indicate the positions of the TATA sequences (T) and the HSEs (1, 2 and 3). Numbers show nucleotide positions relatively to potential transcription start sites. Plasmid DNA (250 μg/ml) from the various constructs was injected into the posterior pole of dechorionated preblastoderm medfly embryos (1–2-h-old) covered with halocarbon oil. Injected embryos were incubated for 22 h at 25 °C in humidified dishes. In heat shock experiments, 20-h-postinjected embryos were incubated for 1 h at 39 °C after which time they were returned to 25 °C for 1 h to recover. In each experiment, groups of 30 well shaped embryos were used to determine CAT activity and plasmid DNA recovery according to Atkinson and O'Brochta (1972). Recovered plasmid DNA was estimated by Southern analysis and this value was used to normalize the CAT data. All activity data shown are expressed as a percentage of the average heat shock activity of C2 construct that was set to 100 and represent mean values ±SE from five independent experiments.

*hsp*70-*cat* construct under both normal and heat shock conditions as described in the figure caption. The results indicate that the 263 bp upstream region of the medfly *hsp*70 gene is sufficient for optimal promoter function at both normal and heat shock conditions. A 5′ deletion that leaves the two HSEs intact, reduces expression levels about four-fold, while a 5′ deletion that leaves only the proximal HSE intact reduces expression levels approximately forty-fold. These data suggest that both HSEs are necessary for the function of the medfly *hsp*70 promoter and that additional upstream sequences enhance promoter activity. Similar results have been reported for the *Drosophila hsp*70 promoter. Functional studies on the *Drosophila* promoter, using germline transformation (Dudler and Travers, 1984) or transfection into cultured cells (Amin et al., 1985), have shown that a region encompassing the two proximal HSEs is sufficient for optimal expression while removal of the 2nd HSE results in a fifty to one hundred times decrease of the heat-inducible expression. The construct C6 that contains a truncated 5′ UTR yielded similar CAT activity to C1 and C2 suggesting that the −263/+105 region of the Cc*hsp*70-B1 gene is sufficient for optimal HSP70 expression at both normal and heat shock conditions. For all constructs, the CAT activity observed in heat

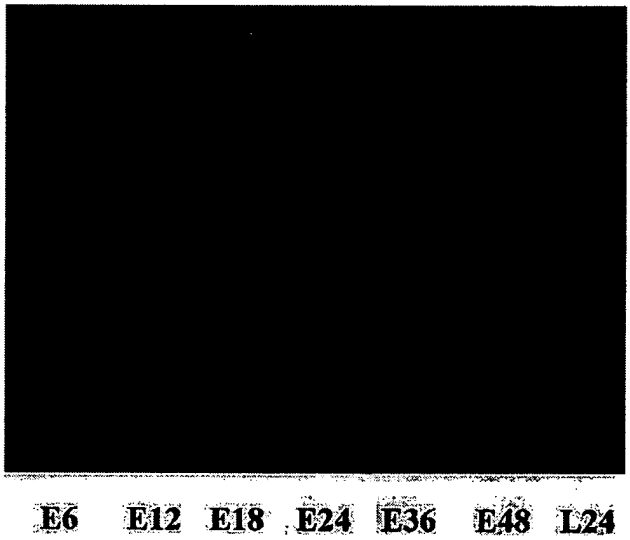


Fig. 3. Developmental expression of CAT in *hsp70-cat* injected medfly emryos. Plasmid DNA from construct C2 (500 μg/ml) was injected into preblasoderm medfly embryos as described in Fig. 2. Thirty embryos or larvae were collected at different times after injections and subjected to CAT assay as described by Atkinson and O'Brochta (1972). Numbers indicate the age of collected embryos (E) and larvae (L) in hours.

shocked embryos was not significantly higher than that of the controls. This is mainly due to the low inducible HSP70 expression in 24-h-old medfly embryos which is approximately forty times lower than that of larvae and adults (unpublished results). The embryonic restriction of HSP70 expression is a widespread phenomenon. Organisms as diverse as fruit flies, sea urchins (Roccheri et al., 1982), frogs (Heikkila et al., 1985) and mice (Morange et al., 1984) restrict HSP70 inducibility in early embryos. Furthermore, the small difference in CAT activity between heat shocked and control embryos could also be due to an overestimation of the constitutive activity in control embryos because of experimental manipulations. As reported for *Drosophila* (Eberlein and Mitchell, 1987; Atkinson and O'Brochta, 1992), the activity observed in control embryos reflects not only the constitutive activity of the *hsp70* promoter, but also the stressed state of the embryos due to dechorionation, desiccation, injection and growth under halocarbon oil. The activity of the *Drosophila hsp70* promoter in medfly embryos was found to be approximately 30% lower than that of the homologous promoter. Germline transformation experiments in medfly showed that the inducible activity of the homologous promoter is at least five times higher than that of the *Drosophila hsp70* promoter (unpublished results). These data are in agreement to those reported by Atkinson and O'Brochta (1992) for *Lucilia cuprina*, and strongly suggest that the *hsp70* and probably other commonly used *D. melanogaster* promoters may not function efficiently in non-drosophilid insects.

4.3. Constitutive promoters

Strong constitutive promoters are important for developing robust transgenic marker systems for detecting, maintaining and recognizing transgenic insects. The *Drosophila actin 5C* gene is actively transcribed in most tissues throughout development and its promoter is widely used for driving constitutive gene expression (Thummel et al., 1988). Comparative studies in transfected *Drosophila* embryos have shown that the *actin 5C* promoter is approximately 10 times stronger than the *hsp70* promoter. However, as with the *hsp70* promoter, the strength of the *actin 5C* promoter was significantly lower in a non-drosophilid insect (Atkinson and O'Brochta, 1992). An actin gene (*CcA1*) has been isolated from medfly but it appears to be muscle-specific and its expression is restricted in late pupal and adult stages (He and Haymer, 1992).

Promoters of constitutive *hsp* genes are also good candidates for driving expression of marker genes. One of the heat shock *Drosophila* cognate genes (*Hsc4*) is expressed abundantly in all developmental stages (Craig et al., 1983). A homolog of this gene has been isolated from medfly (Thanaphum and Haymer, 1998). Functional studies are required to determine the strength and specificity of this promoter. The medfly homolog of the *Drosophila hsp83* gene has also been isolated and shown to be abundantly expressed thought medfly development (unpublished results). Functional analysis of its promoter is currently in progress.

5. Conclusions

During the past six years, combined efforts in several laboratories have led to great progress in the field of gene transfer technology in pest insects. Novel transgenic sexing systems have been developed and proved to work effectively in *D. melanogaster*. However, several components of these systems, such as promoters, regulatory elements and effector genes, may not work efficiently in non-drosophilid insects. Indeed, as pointed out in this review, a number of promoters show different sex and tissue specificity as well as strength between *Drosophila* and other insect species suggesting that homologous components should be considered for developing efficient transgenic sexing and marker systems in pest insects.

Acknowledgements

This work was supported by grants from the Hellenic General Secretariat for Research and Technology and from the FAO/IAEA Division.

References

Alphey, L., Andreasen, M., 2002. Dominant lethality and insect population control. Molec. Biochem. Parasitol. 121, 173–178.

Amin, J., Mestril, R., Lawson, R., Klapper, H., Voellmy, R., 1985. The heat shock consensus sequence is not sufficient for hsp70 gene expression in *Drosophila melanogaster*. Mol. Cell Biol. 5, 197–203.

Amin, J., Anathan, J., Voellmy, R., 1988. Key features of heat shock regulatory elements. Mol. Cell Biol. 8, 3761–3769.

Atkinson, P.W., O'Brochta, D.A., 1992. In vivo expression of two highly conserved *Drosophila* genes in Australian sheep blowfly, *Lucilia cuprina*. Insect Biochem. Mol. Biol. 22, 423–431.

Atkinson, P.W., Pinkerton, A.C., O'Brochta, D.A., 2001. Genetic transformation systems in insects. Annu. Rev. Entomol. 46, 317–346.

Bienz, M., Pelham, H.R.B., 1982. Expression of a *Drosophila* heat-shock protein in *Xenopus* oocytes conserved and divergent regulatory signals. EMBO J 1, 1583–1588.

Berger, E.M., Marino, G., Torrey, D., 1985. Expression of *Drosophila hsp70*-CAT hybrid gene in *Aedes* cells unduced by heat shock. Som. Cell Molec. Gen. 11, 371–377.

Burke, J.E., Ish-Horowicz, D., 1982. Expression of *Drosophila* heat-shock genes is regulated in Rat-1 cells. Nucl. Acids Res. 10, 3821–3830.

Christophides, G.K., Mintzas, A.C., Komitopoulou, K., 2000a. Organization, evolution and expression of a multigene family encoding putative members of the odourant binding protein family in the medfly *Ceratitis capitata*. Insect Mol. Biol. 9, 185–195.

Christophides, G.K., Livadaras, I., Savakis, C., Komitopoulou, K., 2000b. Two medfly promoters that have originated by recent gene

duplication drive distinct sex, tissue and temporal expression patterns. Genetics 156, 173–182.

Christophides, G.K., Savakis, C., Mintzas, A.C., Komitopoulou, K., 2001. Expression and function of the *Drosophila melanogaster* ADH in male *Ceratitis capitata* adults: a potential strategy for medfly genetic sexing based on gene-transfer technology. Insect Mol. Biol. 10, 249–254.

Craig, E.A., Ingolia, T.D., Manseau, L.J., 1983. Expression of *Drosophila* heat-shock cognate genes during heat shock and development. Dev. Biol. 99, 418–426.

Dudler, R., Travers, A.A., 1984. Upstream elements necessary for optimal function of the hsp70 promoter in transformed flies. Cell 38, 391–398.

Eberlein, S., Mitchell, H.K., 1987. Protein synthesis patterns following stage-specific heat shock in early *Drosophila* embryos. Molec. Gen. Genet. 210, 407–412.

Garabedian, M.J., Shepherd, B.M., Wensink, P.C., 1986. A tissue-specific transcription enhancer from the *Drosophila* yolk protein 1 gene. Cell 45, 859–867.

Handler, A.M., McCombs, S.D., Fraser, M.J., Saul, S.H., 1998. The lepidopteran transposon vector, piggy Bac, mediates germ-line transformation in the Mediterranean fruit fly. Proc. Natl Acad. Sci. USA 95, 7520–7525.

Harris, E.J., 1989. Pest status of fruit flies. In: Robinson, A.S., Hooper, G.H. (Eds.), Fruit Flies: Their Biology, Natural Enemies and Control. Elsevier, Amsterdam, pp. 73–81.

He, M., Haymer, D.S., 1992. A muscle-specific actin gene from the Mediterranean fruit fly, *Ceratitis capitata*. Insect Mol. Biol. 1, 15–24.

Heikkila, J.J., Kloc, M., Bury, J., Schultz, G.A., Browder, L.W., 1985. Acquisition of the heat-shock response and thermotolerance during early development of *Xenopus laevis*. Dev. Biol. 107, 483–489.

Heinrich, J.C., Scott, M.J., 2000. A repressible female-specific lethal genetic system for making transgenic insect strains for a sterile-release program. Proc. Natl. Acad. Sci. USA 97, 8229–8232.

Hendrichs, J., Franz, G., Rendon, P., 1995. Increased effectiveness and applicability of the sterile insect technique through male-only release for control of Mediterranean fruit-flies during fruiting seasons. J. Appl. Entomol. 119, 371–377.

Ingolia, T.D., Craig, E.A., Mccarthy, B.J., 1980. Sequence of three copies of the gene for the major *Drosophila* heat shock induced protein and their flanking regions. Cell 21, 669–679.

Katsoris, P.G., Mavroidis, M., Mintzas, A.C., 1990. Identification and characterization of male specific serum proteins in the Mediterranean fruit fly *Ceratitis capitata*. Insect Biochem. 20, 653–657.

Konsolaki, M., Komitopoulou, K., Tolias, P.P., King, D.L., Swimmer, C., Kafatos, F.C., 1990. The chorion genes of the medfly, *Ceratitis capitata*, I: Structural and regulatory conservation of the s36 gene relative to two Drosophila species. Nucleic Acids Res. 18, 1731–1737.

Knipling, E.F., 1955. Possibilities of insect control or eradication through the use of sexually sterile males. J. Econ. Entomol. 48, 459–462.

Krafsur, E.S., 1998. Sterile insect technique for suppressing and eradicating insect populations: 55 years and counting. J. Agric. Entomol. 15, 313–317.

Lindquist, S., Petersen, R., 1990. Selective translation and degradation of heat shock messenger RNAs in Drosophila. Enzyme 44, 147–166.

Lis, J., Castlow, N., de Banzie, J., Knipple, D., O' Connor, D., Sinclair, L., 1982. Transcription and chromatin structure of *Drosophila* heat-shock genes in yeast. In: Schlesinger, J., Ashburner, M., Tissieres, A. (Eds.), A. Heat Schock from Bacteria to Man. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, pp. 57–62.

Logan, S.K., Garabedian, M.J., Wensink, P.C., 1989. DNA regions that regulate the ovarian transcriptional specificity of *Drosophila* yolk protein genes. Genes and Development 3, 1453–1461.

Lombardo, F., Di Cristina, M., Spanos, L., Louis, C., Coluzzi, M., Arca, B., 2000. Promoter sequences of the putative *Anopheles gambiae* apyrase confer salivary gland expression in Drosophila melanogaster. J Biol. Chem. 275, 23861–23868.

Loukeris, T.G., Livadaras, I., Arca, B., Zabalou, S., Savakis, C., 1995. Gene transfer into medfly, *Ceratitis capitata*, using a *Drosophila hydei* transposable element. Science 270, 2002–2007.

Marchini, D., Marri, L., Rosetto, M., Manetti, A.G., Dallai, R., 1997. Presence of antibacterial peptides on the laid egg chorion of the medfly *Ceratitis capitata*. Biochem. Biophys. Res. Commun. 240, 657–663.

Mariani, B.D., Shea, M.J., Conboy, M.J., Conboy, I., King, D.L., Kafatos, F.C., 1996. Analysis of regulatory elements of the developmentally controlled chorion s15 promoter in transgenic *Drosophila*. Dev. Biol. 174, 115–124.

Markaki, M., Craig, K.R., Savakis, C., 2003. Insect population control using female specific pro-drug activation. Insect Biochem. Molec. Biology 34(2), 131–137.

McMahon, A.P., Novak, T.J., Britten, R.J., Davidson, E.H., 1984. Inducible expression of a cloned heat shock fusion gene in sea urchin embryos. Proc. Nat. Acad. Sci. USA 81, 7490–7494.

Mirault, M.-E., Southgate, R., Delwart, E., 1982. Regulation of heat shock genes a DNA sequence up-stream of *Drosphila hsp70* genes is essential for their induction in monkey cells. EMBO J. 1, 1279–1285.

Morange, M., Diu, A., Bensaude, O., Babinet, C., 1984. Altered expression of heat shock proteins in embryonal carcinoma and mouse early embryonic cells. Mol. Cell Biol. 4, 730–735.

Mismer, D., Rubin, G.M., 1987. Analysis of the promoter of the ninaE opsin gene in *Drosophila melanogaster*. Genetics 116, 565–578.

Muller, H.M., Catteruccia, F., Vizioli, J., della Torre, A., Crisanti, A., 1995. Constitutive and blood meal-induced trypsin genes in *Anopheles gambiae*. Exp. Parasitol. 81, 371–385.

Pane, A., Salvemini, M., Delli Bovi, P., Polito, C., Saccone, G., 2002. The transformer gene in *Ceratitis capitata* provides a genetic basis for selecting and remembering the sexual fate. Development 129, 3715–3725.

Papadimitriou, E., Kritikou, D., Mavroidis, M., Zacharopoulou, A., Mintzas, A.C., 1998. The heat shock 70 gene family in the Mediterranean fruit fly *Ceratitis capitata*. Insect Mol. Biol. 7, 1–12.

Parsch, J., Stephan, W., Tanda, S., 1999. A highly conserved sequence in the 3′-untranslated region of the *Drosophila Adh* gene plays a functional role in *Adh* expression. Genetics 151, 667–674.

Parsch, J., Russell, J.A., Beerman, I., Hartl, D.L., Stephan, W., 2000. Deletion of a conserved regulatory element in the *Drosophila Adh* gene leads to increased alcohol dehydrogenase activity but also delays development. Genetics 156, 219–227.

Pelham, H.R.B., 1982. A regulatory upstream promoter element in the *Drosophila hsp70* heat-shock gene. Cell 30, 517–528.

Rina, M.D., Mintzas, A.C., 1988. Biosynthesis and regulation of two-vitellogenins in fat body and ovaries of *Ceratitis capitata* (Diptera). Roux's Arch. Dev. Biol. 197, 167–174.

Rina, M., Savakis, C., 1991. A cluster of vitellogenin genes in the Mediterranean fruit fly *Ceratitis capitata*: sequence and structural conservation in dipteran yolk proteins and their genes. Genetics 127, 769–780.

Robinson, A., 2002. Mutations and their use in insect control. Mutation Research 511, 113–132.

Robinson, A.S., Van Heemert, C., 1981. Genetic sexing in *Drosophila melanogaster* using the alcohol dehydrogenase locus and a Y-linked translocation. Theoret. Appl. Genet. 59, 23–24.

Robinson, A.S., Riva, M.E., Zapater, M., 1986. Genetic sexing in *Ceratitis capitata* using the alcohol dehydrogenase locus. Theoret. Appl. Genet. 72, 453–457.

Robinson, A.S., Franz, G., Fisher, K., 1999. Genetic sexing strains in the medfly, *Ceratitis capitata*: Development, mass rearing and field application. Trends in Entomol. 2, 81–104.

Roccheri, M.C., Sconzo, G., Di Carlo, M., Di Bernardo, M.G., Pirrone, A., Gambino, R., Giudice, G., 1982. Heat-shock proteins in sea urchin embryos: Transcriptional and posttranscriptional regulation. Differentiation 22, 175–178.

Rosetto, M., De Filippis, T., Manetti, A.G., Marchini, D., Baldari, C.T., Dallai, R., 1997. The genes encoding the antibacterial sex-specific peptides ceratotoxins are clustered in the genome of the medfly *Ceratitis capitata*. Insect Biochem. Mol. Biol. 27, 1039–1046.

Rosetto, M., de Filippis, T., Mandrioli, M., Zacharopoulou, A., Gourzi, P., Manetti, A.G., Marchini, D., Dallai, R., 2000. Ceratotoxins: female-specific X-linked genes from the medfly, *Ceratitis capitata*. Genome 43, 707–711.

Ronaldson, E., Bownes, M., 1995. Two independent *cis*-acting elements regulate the sex- and tissue-specific expression of *yp3* in *Drosophila melanogaster*. Genet. Res. Camb. 66, 9–17.

Schutt, C., Nothiger, R., 2000. Structure, function and evolution of sex-determining systems in Dipteran insects. Development 127, 667–677.

Simon, J.A., Sutton, G.A., Lobell, R.B., Glaser, R.L., Lis, J.T., 1985. Determinants of heat shock-induced chromosome puffing. Cell 40, 805–817.

Skavdis, G., Siden-Kiamos, I., Muller, H.M., Crisanti, A., Louis, C., 1996. Conserved function of *anopheles gambiae* midgut-specific promoters in the fruitfly. EMBO J. 15, 344–350.

Swimmer, C., Fenerjian, M.G., Martinez-Cruzado, J.C., Kafatos, F.C., 1990. Evolution of the autosomal chorion cluster in *Drosophila*. III. Comparison of the s18 gene in evolutionarily distant species and heterospecific control of chorion gene amplification. J. Mol. Biol. 215, 225–235.

Swimmer, C., Kashevsky, H., Mao, G., Kafatos, F.C., 1992. Positive and negative DNA elements of the *Drosophila grimshawi s18* chorion gene assayed in *Drosophila melanogaster*. Dev. Biol. 152, 103–112.

Thanaphum, S., Haymer, D.S., 1998. A member of the hsp70 gene family from the Mediterranean fruit fly, *Ceratitis capitata*. Insect Mol. Biol. 7, 63–72.

Thomas, D.D., Donnelly, C.A., Wood, R.J., Alphey, L.S., 2000. Insect population control using a dominant, repressible, lethal genetic system. Science 287, 2474–2476.

Thummel, C.S., Boulet, A.M., Lipshitz, H.D., 1988. Vectors for Drosophila P-element-mediated transformation and tissue culture transfection. Gene 74, 445–456.

Thymianou, S.P., Chrysanthis, G., Petropoulou, K.A., Mintzas, A.C., 1995. Developmentally regulated biosynthesis of two male specific serum polypeptides in the fat body of the medfly *Ceratitis capitata*. Insect Biochem. Mol. Biol. 25, 915–920.

Thymianou, S., Mavroidis, M., Kokolakis, G., Komitopoulou, K., Zacharopoulou, A., Mintzas, A.C., 1998. Cloning and characterization of a cDNA clone encoding a male-specific serum protein of the Mediterranean fruit fly, *Ceratitis capitata*, with sequence similarity to odourant-binding proteins. Insect Mol. Biol. 7, 345–353.

Tolias, P.P., Konsolaki, M., Komitopoulou, K., Kafatos, F.C., 1990. The chorion genes of the medfly, *Ceratitis capitata*. II. Characterization of three novel cDNA clones obtained by differential screening of an ovarian library. Dev. Biol. 140, 105–112.

Tolias, P.P., Konsolaki, M., Halfon, M.S., Stroumbakis, N.D., Kafatos, F.C., 1993. Elements controlling follicular expression of the s36 chorion gene during *Drosophila* oogenesis. Mol. Cell Biol. 13, 5898–5906.

Vlachou, D., Konsolaki, M., Tolias, P.P., Kafatos, F.C., Komitopoulou, K., 1997. The autosomal chorion locus of the medfly *Ceratitis capitata*. I. Conserved synteny, amplification and tissue specificity but sequence divergence and altered temporal regulation. Genetics 147, 1829–1842.

Vlachou, D., Komitopoulou, K., 2001. The chorion genes of the medfly. II. DNA sequence evolution of the autosomal chorion genes s18, s15, s19 and s16 in Diptera. Gene 270, 41–52.

Voellmy, R., Rungger, D., 1982. Transcription of a *Drosophila* heat shock gene is heat-induced in *Xenopus* oocytes. Proc. Natl Acad. Sci. USA 79, 1776–1780.